Project no. 269977

APARSEN
Alliance for Permanent Access to the Records of Science Network

Instrument: Network of Excellence

Thematic Priority: ICT 6-4.1 – Digital Libraries and Digital Preservation

D25.1 Interoperability Objectives and Approaches

Document identifier:	**APARSEN**-REP-D25_1-01-1_7
Due Date:	28-02-2013
Submission Date:	28-02-2013
Work package:	WP25
Partners:	UNITN, FORTH, ESA, FRD, DANS, KB, STFC, CSC
WP Lead Partner:	FORTH
Document status	Final
URN	urn:nbn:de:101-20140516170

Abstract:

This deliverable aims at collecting interoperability objectives from various initiatives and partners and will produce a matrix of solutions and guidelines that can guide the reader to the multi-dimensional and complex landscape of digital preservation interoperability.

The key results reported in this deliverable are:

- An overview of the current projects and initiatives on interoperability in different areas of digital preservation.
- A description of the main interoperability scenarios and challenges encountered by partners and other stakeholders in their daily life activity that served to drive the definition of the main common interoperability objectives and guidelines for digital preservation.
- A detailed analysis of the current solutions adopted to enable semantic interoperability in the domain of Earth Science.
- An analysis of the key questions about global semantic interoperability in digital preservation enabled by the Semantic Web initiative and Linked Data, including an overview of the main strengths and weaknesses of the approach.
- A broad matrix of models, standards and services for interoperability that cover the main areas of digital preservation which can be used as a tool to navigate the complex ecosystem of the current interoperability solutions.
- A gap analysis, which identifies the gaps between the current situation and future interoperability requirements that must be met and proposes possible strategies to fill these gaps.
- A list of recommendations and guidelines for ensuring interoperable digital preservation services.

Delivery Type	REPORT
Author(s)	Barbara Bazzanella (UNITN), Yannis Tzitzikas (FORTH). Veronica Guidetti (ESA), Andrea Della Vecchia (ESA), Emanuele Bellini (FRD), Barbara Sierman (KB), David Giaretta (STFC), Simon Lambert (STFC), Juha Lehtonen (CSC), René van Horik (DANS).
Approval	
Summary	
Keyword List	
Availability	☒ Public

Document Status Sheet

Issue	Date	Comment	Author
0.1	1/06/2012	Outline	Yannis Tzitzikas (FORTH)
0.2	12/07/2012	Addition of possible assignments regarding the exploitation of the results of completed/ongoing APARSEN WPs.	Yannis Tzitzikas (FORTH)
0.3	9/10/2012	Defining the overall structure of the document and addition of possible assignments for contributions by partners	Barbara Bazzanella (UNITN)
0.4	19/10/2012	Editing the initial contribution from UNITN on different sections	Barbara Bazzanella (UNITN)
0.5	5/11/2012	Integrating the contribution from ESA about the results of the task 2530	ESA (A Della Vecchia, V Guidetti)
0.6	20/11/2012	Integrating the contribution from wp25 partners on interoperability initiatives, services, scenarios and challenges	Barbara Bazzanella (UNITN)
0.7	10/12/2012	Integrating the contribution from DANS	René van Horik (DANS)
0.8	14/12/2012	Integrating the contribution from ESA, UNITN and FORTH about evaluation of the scenarios	Barbara Bazzanella (UNITN), Andrea Della Vecchia (ESA), Yannis Tzitzikas (FORTH)
0.9	18/12/2012	Integrating the contribution from FRD	Emanuele Bellini (FRD)

1.0	27/12/2012	Revision of the deliverable based on the internal review by Simon Lambert (STFC)	Barbara Bazzanella (UNITN)
1.1	07/01/2013	Integrating the contribution from FORTH	Barbara Bazzanella (UNITN)
1.2	31/01/2013	Integrating the KB contribution to the scenarios and challenges section	Barbara Bazzanella (UNITN)
1.3	04/01/2012	Revision and refinement of the entire document, improvements to the graphical representations and classification of services.	Barbara Bazzanella (UNITN)
1.4	05/02/2013	Improving the document following the DANS feedback	Barbara Bazzanella (UNITN)
1.5	06/02/2013	Pre-final version	Barbara Bazzanella (UNITN)
1.6	25/02/2013	Internal review	Simon Lambert (STFC)
1.7	26/02/2013	Final version ready for submission	Barbara Bazzanella (UNITN)

Project information

Project acronym:	**APARSEN**
Project full title:	**Alliance for Permanent Access to the Records of Science Network**
Proposal/Contract no.:	**269977**

Project coordinator: Simon Lambert/David Giaretta

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	simon.lambert@stfc.ac.uk / david.giaretta@stfc.ac.uk

Contributors

Participant Number	Participant Short Name	Persons	Comments
25	UNITN	Barbara Bazzanella	Lead Beneficiary
17	FORTH	Yannis Tzitzikas	
11	ESA	Veronica Guidetti; Andrea Della Vecchia	
23	FRD	Emanuele Bellini	
13	KB	Barbara Sierman	
1	STFC	David Giaretta; Simon Lambert	STFC has been replaced as the lead on D25.1 by UNITN
6	CSC	Juha Lehtonen	
12	DANS	René van Horik	

CONTENT

Executive Summary

In recent years, the rising growth of digital information, which needs to be preserved, exchanged and reused in the long term across systems, institutions and platforms, pose new interoperability challenges for digital preservation. A number of initiatives have started to focus on the definition of requirements, technological solutions and best practices in order to define digital preservation interoperability frameworks, services and standards for effectively and reliably access the preserved digital content between interoperable digital preservation systems. This confirms the general agreement that without a broad international consensus on interoperability and appropriately designed technological infrastructures to enable it, digital preservation can not be very effective due to the fragmentation of the underlying ecosystem.

However, interoperability in digital preservation is a very complex issue, which encompasses a multidimensional spectrum of aspects ranging from more technological aspects to include several dimensions of the digital preservation universe (e.g. users, policies, rights), involving different layers of interoperability (e.g. syntactic, semantic, organizational) and different objects (e.g. metadata, Persistent Identifiers, policies) and cutting across different areas of digital preservation (e.g. Provenance, Digital Preservation Services, Persistent Identifiers). Moreover different stakeholder communities deal with a broad range of interoperability challenges and barriers, which affect in many ways different local functionalities and systems.

This deliverable aims at collecting interoperability objectives from various initiatives and partners and will produce a matrix of solutions and provide recommendations that can guide the reader to the multi-dimensional and complex landscape of digital preservation interoperability.

To provide a diagnosis of this ecosystem, an overview of the current projects and initiatives on interoperability in different areas of digital preservation has been conducted, by providing information about more than 60 projects and initiatives. Thanks to the experience of ESA, special focus has been given to projects and initiatives in the Earth Science domain that allowed to investigate semantic interoperability issues and challenges which are common to many other domains, such as ontology mapping, vocabulary alignment, multi-domain thesauri and vocabularies, metadata sharing. The deliverable also includes an analysis of the key questions about global semantic interoperability in digital preservation enabled by the Semantic Web initiative and Linked Data, including an overview of the main strengths and weaknesses of the approach, with a special focus on Cool URIs as an alternative solution to persistent identifiers for digital objects, authors and other entities relevant for digital preservation.

Then the deliverable focuses on the main challenges encountered by partners and other stakeholders in their daily life activity by providing 13 interoperability scenarios evaluated according to three dimensions to provide an indication of the level of relevance of the challenges.

Based on the above, a matrix of models, standards and services for interoperability that cross the main areas of digital preservation which can be used as a tool to navigate the complex ecosystem of the current interoperability solutions was formulated. In future, this tool can be made publicly available as an online search tool, which could be collaboratively extended in the course of the project and even after its conclusion.

Based on the objectives and challenges addressed by the current solutions, and using feedback from the partners, the deliverable identified the main gaps between the current situation and future interoperability requirements that must be met and proposed possible strategies to fill these gaps.

Finally, the deliverable provides a list of recommendations and guidelines for ensuring interoperable digital preservation services.

The reported results provide valuable knowledge and instruments for the VCoE both for raising a common awareness and understanding about interoperability opportunities and challenges in digital preservation and for defining a digital preservation agenda ensuring a coordinated and interoperable

digital preservation ecosystem. In particular the document is relevant for the following aspects for the realization of a VCoE:

- Providing preliminary updated information about the current projects, initiatives, models, standards and services for interoperability in Europe and worldwide.
- Providing a set of recommendations and guidelines on interoperability solutions for specific challenges and areas of applicability.
- Presentation of interoperability best practices in specific domains, which can inspire solutions for other domains and promote cross-fertilization among disciplines and the use of common standards.
- Raising awareness about relevant standardization activities among the relevant user-communities and facilitate the emergence of a diverse offer of services for interoperability in several areas of digital preservation and for different digital resources and related entities.

1 INTRODUCTION

In general, interoperability refers to the ability of two or more systems to exchange information and use the exchanged information. In recent years has become clear that interoperability is one of the key challenges for long-term distributed access to and use of digital resources. Interoperability has been also recognized as one of the potential threats to digital preservation (see for example D3.6 of Parse.Insight project (van der Hoeven, 2010)), since without a broad international consensus on interoperability and appropriately designed technological infrastructure to enable it, digital preservation can not be very effective.

A first step towards an agreement on the main interoperability issues for digital preservation is to define a set of core interoperability objectives and recommendations agreed by the main stakeholders of the services for digital data preservation. This is not a trivial task for many reasons.

First of all, interoperability is a very broad and complex concept, which is conceived on different levels of abstraction ranging from syntactic to semantic interoperability passing through technical, functional and pragmatic perspectives and dealing with many interoperability objects (e.g. metadata, Persistent Identifiers, policies). Secondly, several interoperability issues cut across different areas of digital preservation (e.g. Persistent Identifiers, Authenticity and Provenance, Data Management policies, Storage solutions, Data Quality approaches and so on) showing a very complex and fragmented landscape where there is relatively little harmonization of models, standards and services used in the creation, management and preservation of digital cultural contents. This results in high efforts and costs to guarantee accessibility and long-term reusability of resources across systems and other boundaries. Finally, different stakeholder communities deal with a broad range of interoperability challenges and barriers, which affect in many ways different local functionalities and systems.

We believe that the APARSEN VCoE may provide the right network to start to defragment the interoperability ecosystem in digital preservation and forge strong alliances with major stakeholders to harness global expertise and maximize opportunities.

The main goal of the present document is to report the preliminary results of this effort to capitalize key achievements, address the main open issues, capture best practices, promote shared models and standards and produce guidelines for the development of future interoperability infrastructures, virtualization models and services.

The document starts with a description of the deliverable objectives and its relationships with the tasks 2510 and 2530, explaining the methodology adopted and the role of a Virtual Centre of Excellence for interoperability. Section 2 provides an overview about the definitions of the interoperability concept and layers of interoperability and explains why interoperability is a key issue for digital preservation. Then, follows an extensive description about the main interoperability initiatives and projects related to digital preservation to achieve interoperability at European and worldwide level. A special attention is dedicated to semantic interoperability initiatives in the domain of Earth Science due to the work in the task 2530. Interoperability scenarios and related challenges are presented in Section 3. In Section 4 we connect the interoperability challenges reported by partners (especially as part of the task 2510 activities) regarding different areas of digital preservation to the main models, standards and solutions adopted to address these challenges. In Section 5, we will discuss the gaps between the current situation and the agreed interoperability objectives for digital preservation, proposing guidelines, recommendations and possible solutions (virtualization models, standards and services) to address the main interoperability challenges. A more detailed analysis will be dedicated to the activities related to semantic interoperability within the Earth Science domain, which represented the focus of the task 2530. Finally, we conclude the document providing an overview of the main results of the deliverable and discussing how they can be taken into effective planning for the future. An overview of the deliverable structure is presented in Figure 1.

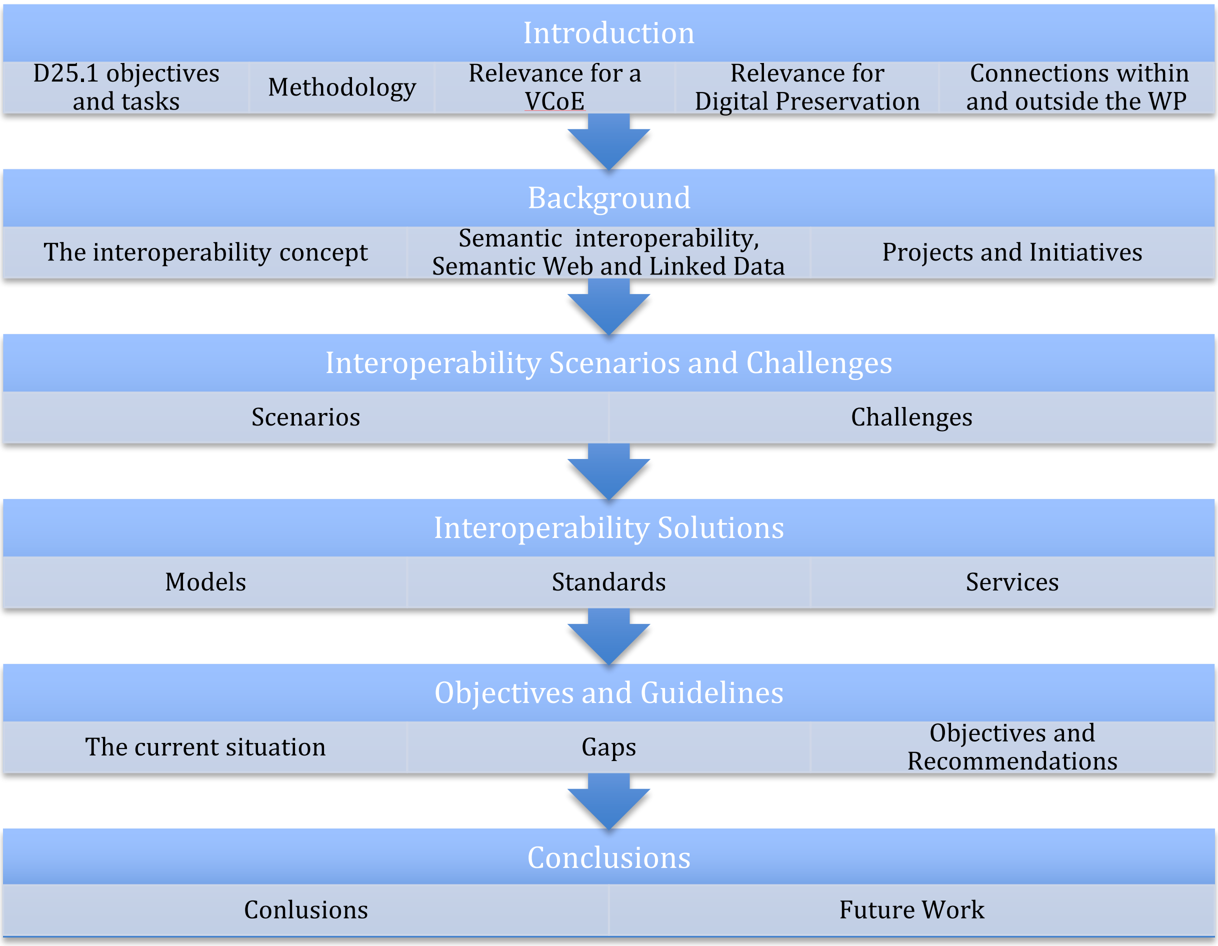


Figure 1: D25.1 Overview

1.1 DESCRIPTION OF THE DELIVERABLE (FROM DOW)

The following description of the deliverable, and related objectives, are taken from the WP25 description in the Description of Work (DoW) document.

> This document will gather the interoperability objectives and guidelines agreed by experts and stakeholders from the participating communities. Subsequently it will propose a useful map/framework/matrix to structure the complex ecosystem of interoperability issues in digital preservation, helping users and key stakeholders to solve their practical interoperability issues (i.e. finding suitable solutions) in different areas of digital preservation and for different interoperability objects.

This deliverable will be the result of Tasks 2510 and 2530.

1.2 RELATED TASKS

This deliverable is related to the following two tasks:

Task 2510: Research and Development of Common Services and Models to Support Interoperability (leader UNITN)

In this task we will gather the conceptual models, services and formats that are used by the partners to address concrete interoperability challenges in digital preservation and try to structure the complex landscape of interoperability initiatives, models, virtualisation of data, and solutions to identify the future objectives and propose the possible recommendations to fill the identified discrepancies and gaps. The identified gaps could be used as input for activities that could lead to new standards and interoperability services. The specific objectives of this task are: 1) to describe ongoing and past projects and initiatives on interoperability in different areas of digital preservation and for different stakeholders and domains; 2) to gather models, standards and services adopted to address different digital preservation interoperability issues in order to provide a concrete tool to classify these solutions and allow a better understanding of this complex ecosystem; 3) to analyze some example interoperability scenarios and challenges encountered by partners and other stakeholders in the domain of digital preservation that serve to drive the definition of the main common interoperability objectives and guidelines for digital preservation. 4) to provide a set of recommendations to fill the gaps and discrepancies between the current situation and the future goals and objectives.

Task 2530: Semantic Interoperability and Scientific Data (leader: ESA)

This task will serve two purposes:

1) to highlight the barriers which currently hamper semantic exploitation of scientific data, with focus on Earth sciences domain;

2) to identify a number of solutions enabling semantic interoperability and exploitation of those data, promoted by the European Space Agency (ESA).

Results of such investigation work will be collected into an internal working document “Semantic interoperability and scientific data in the domain of Earth sciences” (ESA, M22). The overarching goal of this internal document is to allow deriving recommendations with respect to exploiting aspects of scientific data in support of their long term preservation. The document is concepted to provide D25.1 with a comprehensive description of main interoperability issues of one specific scientific domain, and so be instrumental for deriving D25.1 conclusions and recommendations.

The description of the Task 2530 highlights the connections with the Task 2510 and how its results can be integrated into the D25.1. In particular the activity about common models and services to support interoperability, provides a significant contextualization of semantic interoperability issues, objectives and solutions within a specific domain of science.

By describing the main projects and activities related to semantic interoperability, ontologies and knowledge bases in EO domain (reported in Section 2.4.7), Task 2530 identifies a set of concrete interoperability use cases, needs and adopted solutions to support the semantic discovery of EO resources, their preservation and access, and so complements the work of Task 2510. On the other hand the task illustrates also a number of interoperability dependencies, which will be investigated in the D25.2.

The output of the task 2530 is therefore at the base for the definition of objectives, services, interfaces and standards to define a high-level interoperability framework, which will be described in this work and further investigated in the second deliverable of the WP.

1.3 FOCUS AND METHODOLOGY

Digital preservation has been termed as "interoperability with the future" indicating a recent attitude to approach digital preservation as a problem of interoperability.

Following this approach we first aim to provide an overview about on-going and past projects and initiatives on interoperability in different areas of digital preservation and for different stakeholders and domains. Secondly, we will focus on models, standards and services adopted to address different digital preservation interoperability issues in order to provide a concrete tool to classify these solutions and allow a better understanding of this complex ecosystem. A third aspect of our analysis deals with interoperability scenarios and challenges encountered by partners and other stakeholders in the domain of digital preservation that served to drive the definition of the main common interoperability objectives and guidelines for digital preservation. Building on the experience of some work package partners in specific domains, we will analyse some issues in more details. In particular, regarding semantic interoperability we will address solutions and issues related to the use of ontologies for semantic access to Earth Science resources, as investigated at the European Space Agency (ESA). Considering the focus that recent developments of digital preservation have shown on the opportunities offered by the Linked Data approach for data search and integration, a second detailed analysis will focus on global semantic interoperability enabled by the semantic web technologies and the potential lines of collaboration and reciprocal improvement between the Linked Data and Digital Preservation communities. Based on the analysis of past and on-going projects and initiatives, concrete challenges experienced by the main stakeholders and adopted solutions, we aim to investigate the gaps between the current situation and the objectives for the future. The purpose of this analysis is to propose recommendations and guidelines to fill the identified gaps. The work described in the deliverable is also aimed to provide the context for the activities within the Task 2520, by identifying a number of interoperability dependencies, which should be modelled and addressed by the proposed methodology.

Figure 2: D25.1 Focus and Methodology

Work plan

- Literature review and state-of-art analysis on initiatives, models and standards for interoperability. We aim to use the following strategies to collect relevant information on this topic:

  - Exploiting internal and past deliverables (e.g. Task2530 Internal Deliverable, D13.1 on coordination of common standards; D24.1 on provenance interoperability and mapping, D22.1 on Persistent Identifiers and Citability) within the APARSEN project.
  - Exploiting results of other projects and initiatives ((e.g. PARSE.Insight, CASPAR, DIGOIDUNA study).
    - Contacting/interviewing partners about models, standards and services for interoperability, identifying interoperability objectives and needs (gap analysis). Very useful is the contribution from ESA in the Earth Observation (EO) domain (semantic interoperability through thesauri, ontology and their mapping for accessing EO resources).
    - Organizing the collected information into an interoperability reference matrix, which crosses interoperability needs/challenges, objects and solutions with the main digital preservation areas studied in the APARSEN project.
  - Interoperability scenarios and use cases: the idea is to group them around the previously defined digital preservation categories.
  - Identification of common services and initial definition/design of interoperability services for the future.
  - Connecting interoperability solutions (models, standards, services) with challenges and interoperability objects to represent the interoperability landscape.
  - Integration of the task 2510 output with that of task 2530 (led by ESA) on semantic interoperability in the domain of Earth Science into the D25.1 on Interoperability Objectives and Approaches.
  - Identification and analysis of 1) common needs/objectives for interoperability and 2) gaps between the current situation and the goals for the future.
  - Definition of guidelines and recommendations for interoperability in digital preservation.

1.4 THE ROLE OF A VCOE FOR INTEROPERABILITY

As stated in the introduction, the VCoE of APARSEN may play a crucial role in providing expertise, knowledge and capabilities, which are required to create a shared vision about common models, standards and services to enable interoperability in digital preservation.

In this section we will describe how the VCoE can contribute to support interoperability in different areas of digital preservation and how the results of the present work can be useful for this purpose.

The following activities with respect to the realisation of a VCoE are carried out:

- Providing preliminary updated information about the current projects, initiatives, models, standards and services for interoperability in Europe and worldwide. The information collected in this document could be published on a dedicated Web page, which could be used in turn as a collaborative tool to update the content over time.

- Providing a set of recommendations and guidelines on interoperability solutions for specific challenges and areas of applicability.
- Presentation of interoperability best practices in specific domains, which can inspire solutions for other domains and promote cross-fertilization among disciplines and the use of common standards.
- Raise awareness about relevant standardization activities among the relevant user-communities and facilitate the emergence of a diverse offer of services for interoperability in several areas of digital preservation and for different digital resources and related entities.
- Start to define agreed virtualization interfaces supporting access and reuse of digital resources across time, systems and other boundaries (political, economical, national, organizational, disciplinary and so on).
- Contribute to the effort of several APARSEN WPs to identify a common set of terms, whose definition is agreed and shared within the VCoE to improve the understanding and consistency of all the documents produced by the project and introduce a standard terminology to define and design common solutions by key stakeholders. In this respect a common glossary of terms represents a first relevant result for enabling advanced solutions for interoperability for the benefit of the VCoE.

1.5 WHY INTEROPERABILITY IS IMPORTANT FOR DIGITAL PRESERVATION

In this section we discuss the beneficial impact of interoperability for digital preservation and why addressing digital preservation issues with a focus on interoperability may offer significant advantages over current practices for ensuring access, exchange and reuse of digital content in the long term.

- Digital preservation (DP) certainly requires preserving the bits of the digital objects, but this is probably the less difficult task. The preservation of their *accessibility*, *intelligibility*, *provenance*, *authenticity*, *quality (*and many others, e.g. *citability*, *searchability*, etc) is a more complex task. All these requirements can be considered as *interoperability aspects*, in the sense that they can be considered as abilities to apply (now and in the future) successfully in different objects the *same operations* for accessing them, understanding them, rendering them, getting their provenance information, etc. This is why digital preservation has been termed "***interoperability with the future***". Moreover, and as it will be discussed in detail at Section 2, interoperability usually refers to the ability to "***exchange and use*** *information/knowledge between independent systems*" or to "*use the exchanged information in meaningful ways and* ***without special effort*****".** As a consequence, achieving interoperability (according to this definition), implies ability to exchange/use information *without special effort*, thus preservation of accessibility, intelligibility, etc, without special effort. Using standardized interoperable solutions ensures that DP system implementation will not depend on a single vendor.
- Expressing crucial digital preservation challenges as interoperability challenges has a beneficial impact not only for the design and implementation of scalable technical solutions, but also for the definition of a common research agenda agreed by stakeholders, which are concerned with long-term preservation and stakeholders that are focused on building interoperable digital environments. By recognizing that common needs and issues are in play, it should be easier to adopt integrated solutions and expand the applicability of standards and models developed within a certain context to data created and used by other communities. In this way, we can expect that systems, which allow to access and reuse information across

technical, organizational, political and social boundaries should also be more effective in enabling the access and reuse of these resources in the long term.

- A study conducted by the EU in 2011 contains on overview of the most important topics for Digital Preservation Research. "Interoperability" is mentioned as one of the ten research topics (See: "Research on Digital Preservation within projects co-funded by the European Union in the ICT programme" by Stephan Strodl, Petar Petrov, Andreas Rauber (May 2011) http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-future-digital-preservation_en.html).

  The study states: "a great deal of formats and format types exist already and new emerge every day. All present solutions specialize only in a subset of these. There are repositories that could handle any kind of digitally encoded data, however organizations and institutions are often forced to use a number of solutions. Therefore, interoperability and trust between these different service providers is essential for digital preservation" (page 25).
- Digital preservation can be conceived as an interoperability exercise along the entire spectrum of steps that form the lifecycle of a digital object—from its creation to its re-use through the process of preservation. A fundamental aspect of this exercise is the adherence to digital preservation standards, as pointed out by (National Information Standards Organization, 2004) "An institution must ensure that its standards are in line with those used across the digital library community to enable interoperability where possible". To this purpose, digital preservation standards should not be conceived from a repository-centric point of view but should be defined as a set of functional requirements which can be implemented by multiple systems with different hardware and software platforms, data structures, and interfaces to manage and exchange data in the medium and long term with minimal loss of content and functionality.
- Digital preservation infrastructures should be designed to be integrated (i.e. to interoperate) with existing systems (e.g. archiving and management systems) in use by the involved institutions and stakeholders. Therefore, a digital preservation infrastructure should be an interoperability infrastructure, which accommodates preservation services that are hosted by third-parties and supports access to remote and distributed third-party preservation action services. Beside the interoperability with external systems and services, the infrastructure should be also internally interoperable. This means that the components of the framework should be optimized to assure that various layers of the framework are fully interoperable.
- Very often, DP has to deal with data (e.g. cultural heritage data) that are syntactically and semantically heterogeneous, multilingual, multicultural, semantically rich and highly interlinked. Making this content mutually interoperable so that in can be searched, accessed and reused in the long time is a big challenge for DP. The problem involves different levels of interoperability. On a syntactic level, it is needed to harmonize different character sets, data formats, notations and collection records adopted in different collections but also to agree on communication protocols for information exchange. Beside syntactic interoperability issues, the preservation of highly heterogeneous content implies a problem of semantic interoperability. Different metadata standards are in use by different institutions to describe the same type of content, metadata formats may be interpreted differently, data is encoded at different levels of precision, vocabularies and ontologies used in describing the content are different and ontology alignment and mapping is hard to completely automate.

  Probably the most challenging interoperability issue is at the organizational level. The multi-organizational nature in which content is collected, maintained and published poses new issue for DP dealing non only with formats and technical standards but also with different policies, rights and restrictions management, mandates, roles and responsibilities.

1.6 ARTICULATION WITH THE REST APARSEN WPS AND TASKS

In this section we first report the connections of the present work to the remaining WP25 activities mainly related to the Task 2520. Secondly, we present the relationships with the work within other WPs in the project.

1.6.1 Connections with the remaining WP 25 activities and tasks

Here we describe the connections between the current deliverable (D25.1) and the D25.2 (due Aug 2013) since both belong to the same WP.

We could say that each interoperability objective/challenge, like those described in the current deliverable, is a kind of demand for the *performability* of a particular *task* (or tasks). In Task 2520 we will identify such tasks, we will reflect on their dependencies and on how these can be modelled (as described in the DoW). An informal map of the task-dependency approach is sketched in the following figure. In brief, for performing one task upon a module we may need to have available other modules (see association *includeAvailabilityOf*) and/or be able to perform other subtasks (see association *implies*). Modules have types, which can be hierarchically organized, while the notion of *profile* can be used for making explicit the assumptions about the designated community knowledge.

The ultimate objective is to propose a modelling approach that enables the desired reasoning, e.g. task performability checking, which in turn could greatly reduce the human effort required for periodically checking or monitoring whether a task on an archived digital object or collection is performable, and consequently whether an interoperability objective is achievable. Such services could also assist preservation planning, especially if *converters* and *emulators* can be modeled and exploited by the dependency services.

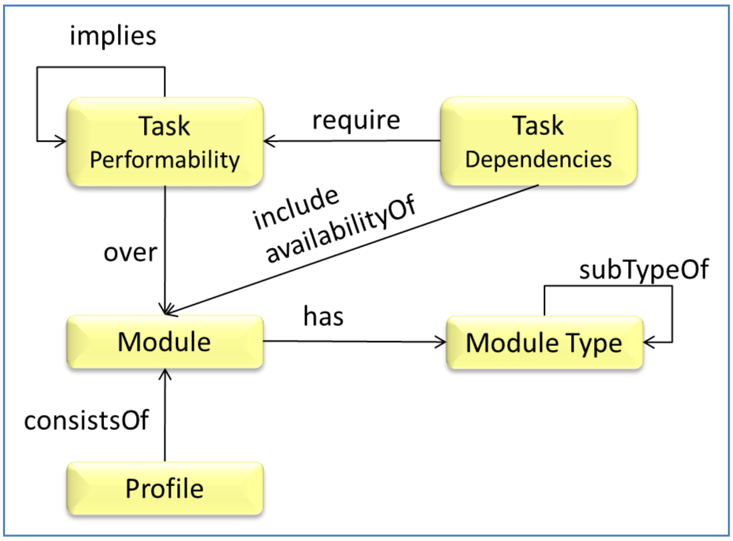


Figure 3: An informal map of some basic notions of the dependency management approach that will be elaborated in the forthcoming D25.2

1.6.2 Connections with the other APARSEN WPs

We have already stressed that interoperability issues and challenges pervade all the main areas of digital preservation. Therefore, it is quite obvious that there are many connections between the present work and the results from other WPs of the APARSEN project.

Here we describe how this work is related with the other work packages and tasks of APARSEN.

This WP will exploit results of several completed, or ongoing WPs, specifically:

Workpackage	***Relevant Connections***
WP24 (Authenticity and Provenance)	The results of the present deliverable are related to the work within the WP24 (especially those reported in the D24.1, *Report on Authenticity and Plan for Interoperable Authenticity Evaluation System*). The proposed model, guidelines and reasoning rules for managing provenance information can be used to ground the interoperability objectives and guidelines presented in the D25.1.
WP26 (Annotation, reputation and Data Quality),	Semantic interoperability and standards for information exchange is one of the issues that connect the results of this deliverable with the WP26 topics. Another relationships is represented by interoperability between identifiers systems which represent a fundamental prerequisite for data quality assessment.
WP22 (Identifiers and Citability)	There is a clear connection with the main goal of the WP22 that is proposing an interoperability framework for PIDs. The use of PIDs and their interoperability is a crucial aspect for allowing interoperability at different levels such as metadata interoperability, knowledge access and reusability, Citability, provenance and data quality assessment.
WP21 (Preservation Services)	There are many connections between the work in the WP21 and the content of the D25.1: • Interoperability is a key issue for the development and integration of preservation services, which allow management, sharing, publication and long-term preservation of distributed data. In particular, if preservation services are to form part of an e-infrastructure for digital preservation, interoperability is a prerequisite to allow these services to integrate effectively. • Interoperability is a basis for services of wide applicability, enabling their use in a diversity of domains and data types. • The emergence of the micro-services approach, that is an approach to digital curation based on devolving curation function into a set of independent, narrowly scoped services, makes the interoperability between these services a key aspect to implement the complex function needed for effective digital curation.
WP11 (Common vision).	The results of the current deliverable can be related with the following candidate objectives for the common vision: • Having a common understanding of the digital

	preservation landscape in particular regarding interoperability gaps, challenges and issues. • Providing a set of recommendations and guidelines on interoperability solutions for specific challenges and areas of applicability. • Presentation of interoperability best practices in specific domains, which can inspire solutions for other domains.
WP13 (Coordination of common Standards).	There are contexts where interoperability can only be achieved through the adoption of common standards and agreed policies. Since WP13 focus on identifying (and in case propose) common standards for enhancing accessibility to information, we envision several connections with this WP. In particular relevant information can be derived from the analysis of current standards in use by partners for their digital preservation activities (Task 1310) and by industry communities (Task 1330). The results can be connected with the interoperability solutions analyzed in this document (Section 4). Interesting insights for defining interoperability objectives and requirements can be found in the context of the task 1320 which aims at identifying new standards requirements. Finally use cases and scenarios in Industrial contexts could be suggested by the results of the Task 1330.
WP14 (Common Testing Environments)	The D25.1 identifies a set of interoperability objectives, scenarios and solutions in the domain of digital preservation, which could be used within the WP14 to design testing procedures within a testing environment.

2 BACKGROUND: THE INTEROPERABILITY CONCEPT

Interoperability is one of the key pillars of the European Commission Digital Agenda for Europe[1] and has become a critical issue in distributed and heterogeneous digital preservation environments where autonomous systems should co-operate and rely on each other to provide integrated and cross-boundary services.

Several definitions have been proposed for the term interoperability and many attempts have been made to identifying different levels of agreement at which interoperability can be realized. Therefore, definitions of interoperability range from very generic conceptualizations to very specific and technical conceptualizations that apply to specific systems or areas of investigations. We start to discuss some of these definitions in order to identify the level of analysis adopted in the present work.

According to the IEEE Standard Computer Dictionary, interoperability refers to "*the ability of two or more systems or components to* ***exchange information and to use the information*** *that has been exchanged*" (Geraci, 1991). The general idea that interoperability is a matter of exchanging information and using this information properly can be found in (Heiler, 1995), which defines interoperability among components of large-scale, distributed systems as "*the ability to exchange services and data with one another. It is based on agreements between requesters and providers on, for example, message passing protocols, procedure names, error codes, and argument types*".

[1] http://ec.europa.eu/information_society/newsroom/cf/digital-agenda.cfm (see pillar II, Interoperability and standards)

A similar definition is suggested by Taylor (Taylor, 2004), which defines interoperability as "*the compatibility of two or more systems such that they can exchange information and data and can use the exchanged information and data **without any special manipulation***".

The last definition introduces the idea that interoperability is not only the ability of multiple systems to exchange and use data, but also the ability of exchanging data without any specific effort for transforming the data that have to be exchanged. Similarly, in (ALCTS/CCS Commitee on Cataloging: Description and Access, 2000) interoperability is defined as "*the ability of two or more systems or components to exchange information and use the exchanged information **without special effort on either system***". In his paper on identifier interoperability (Paskin, 2006), Norman Paskin adds the idea that interoperability is the ability of independent loosely-coupled systems to exchange meaningful information, collaborate and communicate and initiate actions from each other, in order to operate together to **mutual benefit**. The notion that interoperability is a kind of interaction, which allows the interoperating systems to achieve **common goals** for their mutual benefit, can be found in the definition of the European Interoperability Framework[2]: *"Interoperability, within the context of European public service delivery, is the ability of disparate and diverse organisations to interact towards mutually beneficial and agreed common goals, involving the sharing of information and knowledge between the organisations, through the business processes they support, by means of the exchange of data between their respective ICT systems".*

In summary, three main aspects of interoperability emerge from the previous definitions.

1. to **exchange and use** information/knowledge between independent systems;
2. to use the exchanged information in meaningful ways and **without special effort**;
3. to use the exchanged information to achieve **common goals for the mutual benefit** of the interoperating systems.

Various aspects or layers of interoperability have been identified in literature. One of the first distinctions that have been introduced is between **syntactic** and **semantic interoperability** (Park & Ram, 2004).

Syntactic interoperability. If two or more systems are capable of exchanging data, they are exhibiting syntactic interoperability, which is required for any attempt of further interoperability. Syntactic interoperability requires agreement on the structure of data (i.e. data formats) and communication protocols. MARC, XML or SQL are examples of standards that provide syntactic interoperability. Syntactic interoperability deals also with lower-level data formats, such as ensuring alphabetical characters are stored in ASCII format in both of the interoperating systems.

It is worth to notice that the use of the same data formats and protocols does not guarantee that two systems can interoperate at syntactic level. This is because the same structural standard and rules can present subtle differences from system to system or from network to network (think for example of the MARC standard dialects).

Semantic interoperability. Beyond the ability of two or more computer systems to exchange information, semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems. Going back to the Heiler's definition, semantic interoperability ensures that the exchanges between requesters and providers make sense, that means that the requester and the provider have the same understanding of the meanings of the requested service and data. According to (Park & Ram, 2004) "semantic interoperability is the knowledge-level interoperability that provides cooperating businesses with the ability to bridge semantic conflicts arising from differences in implicit

[2] http://ec.europa.eu/isa/documents/eif_brochure_2011.pdf

meanings, perspectives, and assumptions, thus creating a semantically compatible information environment based on the agreed concepts between different business entities".

To achieve semantic interoperability, both sides must defer to a common information exchange reference model. The content of the information exchange requests should be unambiguously defined: what is sent is the same as what is understood.

The European Interoperability Framework[3] identifies and defines three dimensions of interoperability **organisational**, **semantic** and **technical.**

Organizational interoperability. This level of interoperability is concerned with integrating business processes and addressing user requirements "by making services available, easily identifiable, accessible and user-oriented".

Semantic interoperability. This aspect of interoperability is concerned with ensuring that the precise meaning and formats of exchanged information is understood and preserved. Semantic interoperability deals also with addressing the challenges of integrating information across linguistic, cultural, legal and administrative boundaries in a meaningful manner. The multilingual delivery of services to users is an example of semantic interoperability challenge (Miller, 2000).

Technical interoperability. This layer of interoperability is concerned with formalizing technical specifications in order to ensure the functional linking of systems and services. It includes key aspects such as open interfaces, interconnection services, data integration and middleware, data presentation and exchange, accessibility and security services.

Miller (Miller, 2000) has expanded the concept into a broader context proposing several flavours of interoperability.

Technical interoperability. This aspect of interoperability is concerned with developing (and promote the convergence of) communication, transport, storage and representation standards to facilitate the access and the use of contents.

Semantic interoperability. This flavour of interoperability deals with the harmonization of the terms used in different systems to describe resources. Ongoing work focuses on the development and distributed use of thesauri.

Political/Human interoperability. This aspect of interoperability concerns the social environment that is (directly or indirectly) involved in the decisions of making resources more widely available including the organizations, their staff and the end users. Political/human interoperability can be realized though widespread dissemination and training actions, that is essential to ensuring the long-term use of any service.

Inter-community interoperability. Apart from issues related to the manner in which information is described, stored and distributed, there are also issues regarding community and disciplinary boundaries across which resources (and particularly scientific resources) need to be shared. These issues pertain the layer of interoperability, which is called by Miller Inter-community interoperability.

[3] http://ec.europa.eu/idabc/en/document/2319/5644.html

Common solutions at this level are fundamental to the long-term benefit of the sectors concerned and, more importantly, to the benefit of the end-user, who routinely behaves in a cross-disciplinary manner, and is often hampered by institutional barriers.

Legal interoperability. This layer of interoperability is about the legal requirements and implications involved in making resources more widely available. Different systems may operate under different legal regimes, which prevent interoperability, or at least make the legality of interoperability uncertain. Think for example to public sector data which may include personal data usually protected by strict stipulation over their use or certain type of data which are protected by Intellectual Property Rights (IPR).

International interoperability. The last level of interoperability described by Miller includes all the other levels but on an international scale characterized by differences in technical approaches, working and organization practices. Language differences and cultural issues are very important for this layer of interoperability.

According to Miller, in order to be interoperable, "one should actively be engaged in the ongoing process of ensuring the systems, procedures and culture of an organization are managed in such a way as to maximize opportunities for exchange and re-use of information, whether internally or externally". To design an interoperable system, not only technical aspects should be taken into account, but also and semantic, organisational, cultural and legal issues should be taken in strong consideration.

Interoperability is also a major concern in many countries. The e-Government Interoperability Framework (e-GIF) in UK, for example, focuses on the government's technical policies and specifications for achieving interoperability and Information and Communication Technology (ICT) systems coherence across the public sector. The key policies of the framework cover interconnectivity, data integration, content management metadata and e-services access. To be compliant with e-GIF, a system should satisfy two requirements: 1) to be truly interoperable which means to ensure the coherent exchange of information and services between systems, that is the ultimate test for interoperability and 2) it must me possible to replace any component or product used within an interface with another of a similar specification while maintaining the functionality of the systems (CabinetOffice, 2004). Another example, described in more details in Section 2.4.8 is the ISA (Interoperability Solutions for European Public Administrations) Programme, an EC programme to foster interoperability, sharing and re-use between European public administrations.

As emerges from the definitions above, interoperability is a complex and multi-layered concept and a crosscutting concern (Rothenberg, 2008), which encompasses a multidimensional spectrum of aspects ranging from more technological aspects to include several dimensions of the digital preservation universe (e.g. users, policies, disciplines, rights, and many others). Moreover, different communities and disciplines may have very heterogeneous interoperability requirements since their needs with regard to data management and curation vary considerably. Data are homogeneous in some communities, highly heterogeneous in others, and this may have strong impact on the development of interoperable solutions for data exchange (for example based on metadata standards). In some contexts data must be immutable and its provenance and integrity must be ascertained (e.g. astronomy), while in other contexts data can be changed in the course of its lifecycle or must be erasable at any time for legal reasons. Open access policies are in play in some contexts, licensing and strict privacy issues are addressed in others. From these premises, it follows that devising an appropriate solution to the digital preservation interoperability challenges it is far from being a merely technical problem and the diversity of the community requirements makes it impossible to aim for a single strategy or system for economical, political, organizational and social reasons. Interoperability is crucial to address issues like access, provenance, citability, data quality assessment and many others, going far beyond the technical level to embrace a much wider horizon where organizational, social and business strategies must be taken into account in considering effective solutions. If an all-encompassing perspective is taken, including technical, social, geographical, organizational and many other factors, a comprehensive picture of this complex landscape can be provided, enhancing the understanding of its faces and orienting strategies for finding specific solutions. This multidimensional perspective is fundamental to make a diagnosis of the current situation and propose integrated actions towards co-ordinated sustainable and agreed interoperability solutions to promote the potential of digital preservation. A result of this multidimensional analysis can be seen in the matrix of interoperability solutions presented in Section 4.2 where the solutions have been clustered around digital preservation areas and digital preservation objects, and have been linked to interoperability layers.

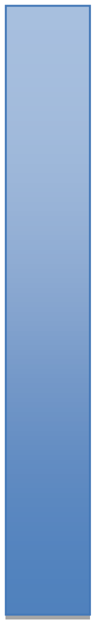


Figure 4: Interoperability Layers

2.1 THE CONCEPT OF INTEROPERABILITY AS A FRAMEWORK FOR DIGITAL PRESERVATION

Recent developments in digital preservation have begun to approach digital preservation as a problem of interoperability. Digital preservation has been termed “interoperability with the future” indicating that digital preservation is concerned with accessing and using information in a useful and meaningful way across time. The idea beyond this approach is resumed in the concept of **temporal interoperability** proposed by Hedstrom (Hedstrom, 2001), which defines temporal interoperability as “the ability of current systems or legacy systems to interoperate with future systems that may use new formats, data models, languages, communication protocols, and hardware.”

However, the techniques used for contemporaneous interoperability are applicable for temporal interoperability (i.e. digital preservation), indicating many potential commonalities and points of synergy between interoperability in real time and digital preservation. These synergies and common objectives can be found at all the levels of interoperability that we have described in the previous section. For example, if you consider the technical level, contemporaneous interoperability and digital preservation share the goal of enabling the access and reuse of information independently from the technological platforms, languages, formats and so on. Therefore solutions for enabling contemporaneous interoperability can be adopted to develop digital preservation strategies. The use of persistent identifiers is an example of a contemporaneous interoperability solution that can have a strong impact in terms of digital preservation.

Another example deals with semantic interoperability where models or conceptual frameworks to support long-term digital preservation may have an impact on contemporaneous interoperability. An example of this is provided by the OAIS reference model, which defines the key concept of Representation Information:

> *"The information that maps a Data Object into more meaningful concepts. An example of Representation Information for a bit sequence which is a FITS file might consist of the FITS standard which defines the format plus a dictionary which defines the meaning in the file of keywords which are not part of the standard. Another example is JPEG software which is used to render a JPEG file; rendering the JPEG file as bits is not very meaningful to humans but the software, which embodies an understanding of the JPEG standard, maps the bits into pixels which can then be rendered as an image for human viewing*."

In addition the CASPAR project identified a number of layers of Representation Information which need to be captured as a piece of digitally encoded information and that is ingested into an archive for preservation (including re-use), (Rep Info in **Figure 5: CASPAR Information Flow Architecture**

Other layers include Access Control and Persistent Preservation Infrastructure, which include for example a Key Store if the data is encrypted.

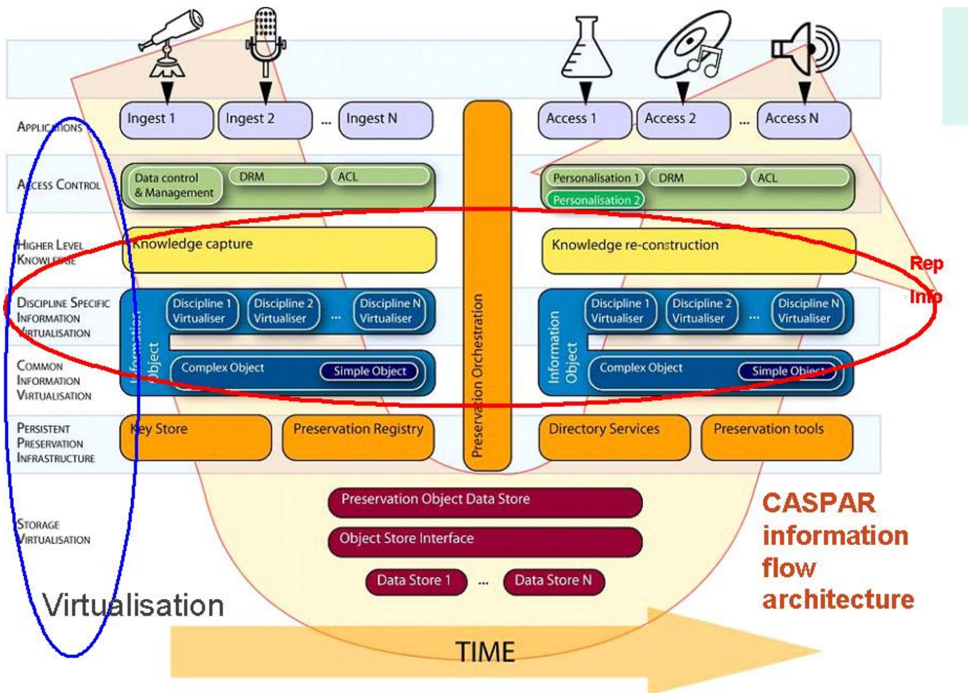


Figure 5: CASPAR Information Flow Architecture

The important point is that at each stage an artefact, usually digital, is created, for example the access control might consist of a description of the rights associated with the object, formalised in some rights description language, and a list of individuals and their roles.

When the preserved information is to be re-used, whether in the future or immediately in a different environment (for example a different discipline or legal jurisdiction), those artefacts must also be re-used, and there may be difficulties in doing this. For example the rights description language may no longer be used, or may not be applicable for some reason. Similarly the encryption algorithm may not be in easily usable. Therefore the same techniques must be applied to these digital artefacts.

This recursion continues for each digital object until, in terms of preservation, the Designated Community can use the encoded information. For contemporaneous interoperability, the same concepts may be applied to guide the repository wishing to make its data more widely usable.

The CASPAR project tried to ensure that the use of the Representation Information is as easy and automated as possible, and is widely usable beyond the Designated Community. This demands increasing automation in the access, interpretation and use of Representation Information, and also the provision of more clues to users from different disciplines.

For the latter one can begin by offering some common views on data – for example allowing easier use in generic applications – by means of virtualisation. An example of this would be where the information embodies an image, to make this fact explicit in the Representation Information so that an application would know that it makes sense to handle the data as a 2-dimensional image. In particular the data can be displayed; it has a size specified as a number of rows and columns. This type of virtualisation is common in many other, non-preservation related, areas. It is the basis on which computer operating systems can work surviving many generations of changes in technologies, on a variety of hardware. For example, the operations that a disk drive must perform can be specified and used throughout the rest of the operating system, but the specifics of how that is implemented are isolated within a driver library. The underlying idea here is to define, in software terms, a set of interfaces, which can be implemented on top of a variety of specific instances, which will change over time.

In CASPAR, virtualization techniques and tools (e.g. Virtualization Manager tool) – i.e. the creation of a software layer on a modern computer system that allow the obsolete systems to execute on a virtual machine - have been proposed as key infrastructure components to support different aspects of the digital preservation of digital objects. Figure 5 shows in more details how CASPAR expects to use virtualisation including:

- Digital Object Storage virtualisation
- Common information virtualisation
- Discipline-specific information virtualisation

Virtualisation also applies to the higher-level knowledge, access control and digital rights management and processes. The Warwick Workshop, Digital Curation and Preservation: “Defining the research agenda for the next decade”, held in November 2005, noted that virtualisation is an underlying theme, with a layering model illustrated as follows:

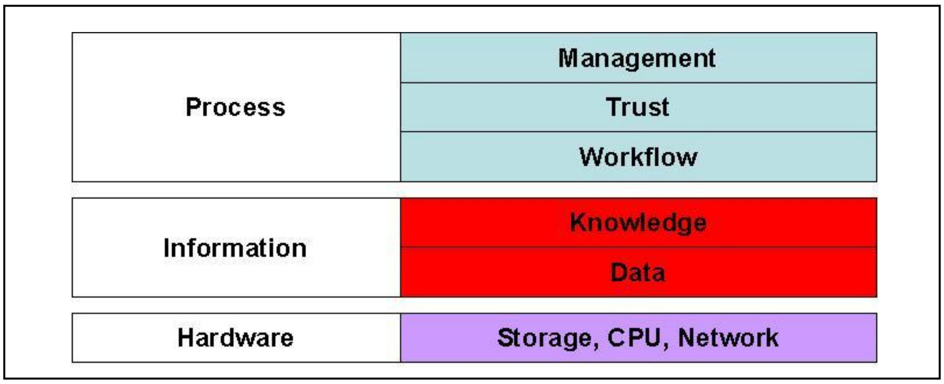


Figure 6: Layers of virtualization (from the Warwick Workshop)

The common research issues identified were:

Automation and Virtualisation	1. Develop language to **describe data policy** demands and processes, together with associated support systems.
	2. Develop **collection oriented description** and transfer techniques.
	3. Develop data description tools and associated generic migration applications to **facilitate automation**.
	4. Develop **standardised intermediate forms** with sets of coder/decoder pairs to and from specific common formats.
	5. Develop **code generation tools** for automatically creating software for format migration.
	6. Develop techniques to allow **data virtualisation of common science objects**, with at least some discipline specific extensions.
	7. **Management and policy specifications** will be need to be formalised and virtualised.
	8. Further **virtualisation of knowledge** – including developments of interoperable and maintainable ontologies.
	9. Develop **automatic processes for metadata extraction**

Specific research topic included:

Virtualisation	10. Continuing work on ways of **describing information all the way from the bits upwards**, in standardised ways – “virtualisation”. Work is needed on each of the identified layers in section A1.2.
	11. **Knowledge virtualization** involving Ontologies and other Semantic Web developments are required to enable the characterization of the applicability of a set of relationships across a set of semantic terms.
	12. Develop use of **data format description languages** to characterize the structures present within a digital record, independently of the original creation application.
	13. It is important to make significant progress on dealing with **dynamic data**

	including databases, and object behaviour.
	14. **Representation Information tools**, probably via layers of virtualisation to allow appropriate normalisation, including mature tools for dealing with dynamic data including databases.
	15. Additional work on **preservation strategies and support tools**, from emulation to virtualisation.
	16. Develop increasingly powerful virtualisation **tools** and techniques, with a particular emphasis on knowledge technologies.
Automation	17. Develop protocols and information management exchange mechanisms, including synchronisation techniques for indices etc., to **support federations.**
	18. **Standardised APIs** for applications and data integration techniques
	19. Fuller development of **workflow systems and process definition** and control.
Support	20. Develop simple semantic **descriptions of Designated Communities**.
	21. **Standardise Registry/Repositories for Representation Information** to facilitate sharing.
	22. Develop **methodologies and services for archiving personal collections** of digital materials.
Hardware	23. Develop and **standardise interfaces** to allow "pluggable" storage hardware systems.
	24. **Standardise archive storage API** i.e. standardised storage virtualisation.
	25. Develop **certification processes for storage systems**.
	26. Undertake research to characterise types of read and **transmission errors** and the development of techniques which detect and potentially correct them.

Table 1: Research Topics (Warwick Workshop)

However, virtualization is not a panacea and, as noted in (Giaretta, The CASPAR Approach to Digital Preservation, 2007), one cannot expect that it can be applied everywhere. Moreover, where virtualization techniques can be adopted, interfaces can become obsolete and require to be re-virtualized. Nevertheless, virtualization is an important concept in considering digital preservation as the effort of ensuring interoperability in the time dimension (i.e. interoperate with the future) since virtualization strategies and technologies contribute to guarantee that data or digital objects remain authentic and accessible over a long period of time, despite the obsolescence of software and systems.

Virtualization is also one of the techniques recommended in the PARSE.Insight project (D2.1, p.15) to enable interoperability of science data infrastructures. By virtualizing access and use of science data infrastructure components, data management operations like, for example migration between preservation environments, are facilitated, promoting interoperability between services, grid operations and existing archive systems. An example can be found in the Earth Science (ES) domain.

Virtualizing access and reuse of scientific data is widely recognized as a key challenge within the Earth Science (ES) domain. There is a need of preserving the rising flood of ES data and more importantly the associated knowledge (e.g. technical and scientific documentation, algorithms, data handling procedures, etc.) to ensure its meaningful long-term accessibility and interoperability between different scientific communities. This deals with the issue of abstraction of information.

The research topics described above illustrate the potential advantage of expressing digital preservation as an interoperability challenge. The general idea is to reduce the incompatibilities between older and new technologies and systems by virtualization rather then to transform old data to achieve the requirements of new systems.

Addressing contemporaneous and long-term interoperability as a unique issue has the potential of adopting an integrated approach to effective digital management and preservation as well.

2.2 SEMANTIC INTEROPERABILITY IN THE DOMAIN OF EARTH SCIENCE

As a case study this section contain a more in depth report on the achievement made concerning semantic interoperability in a specific domain, namely Earth Science.

Interoperability across heterogeneous systems can be hampered by differences in semantics, especially when multiple scientific communities and many different users are involved. Access to (and reuse of) Earth Science resources represent a very interesting example where reconciling these differences, enabling semantic interoperability, has become a crucial issue. In the current work we focus on interoperability of data access, which is the level of interoperability that should be pursued to favour easy and cost-effective access means to heterogeneous EO data, despite the differences between data sets and the accessed systems.

Earth Observation resources, products and services are useful for a whole lot of different applications, ranging from security to climate monitoring and from research to operational activities. Each of these application fields has its very specific knowledge and expertise, which does not necessarily include any understanding of the Earth Observation domain. A typical situation is the one described by the Semantic Interoperability scenario, proposed in Section 3, describing the situation the domain experts, interested in using EO data to support their activities, have to cope with. One way to help them is to set up web services that allow easily identifying and getting access to the EO resources that best fit their needs. In particular, permitting the discovery through controlled terminology the application domain experts are familiar with, and EO resource access through centralised and/or federated catalogues (Heterogeneous Mission Accessibility Cookbook, 2012).

2.3 GLOBAL INTEROPERABILITY, SEMANTIC WEB AND THE LINKED DATA APPROACH

Since the first definition of the Semantic Web by Tim Berners-Lee (Berners-Lee, The Semantic Web, 2001) over ten years ago, there has been a growing interest and application of the W3C[4] standards to enable semantic web-scale interoperability. Recently, digital preservation communities and cultural heritage data providers have also started to manifest a positive attitude toward this approach (Hyvönen, 2012), despite an initial scepticism and some criticisms, and a growing number of initiatives are now adopting the Semantic Web technologies to address some digital preservation issues, like enhanced web searching, data classifications and integration. We will describe some of these initiatives in the course of this document. Here, we briefly present the Semantic Web (and Linked Data) paradigm as a solution for global interoperability.

According to the W3C definition, the Semantic Web is a collaborative effort, which “provides a common framework that allows data to be shared and reused across application, enterprise and community boundaries.” The term was coined by Tim Berners-Lee, the inventor of the WWW, which

[4] http://www.w3.org

in his original 2001 article (Berners-Lee, The Semantic Web, 2001) envisioned the Semantic Web, as an extension of the Web in which information is provided in a well-defined meaning by enabling computers and people to work in cooperation.

The Semantic Web vision is based on the idea to extend the principles of the Web from documents to data by defining and describing the relations among data on the Web. These meaningful relationships can be established between any named resources in the Web documents by enabling an automatic integration of data. The Semantic Web data are therefore characterized by two aspects: 1) an explicitly defined structure (like in databases) and 2) semantics (like in ontologies).

The Semantic Web architecture is illustrated by the Semantic Web Stack, also known as Semantic Web Cake or Semantic Web Layer Cake (see Figure 7). It is a kind of layered architecture for interoperability. Standardization on one level allows standardizing an upper level and so on. For instance standardizing the structure (e.g. XML) allows standardizing structure queries (e.g. XQuery). Such stacks can be conceived as a kind of layered interoperability objectives, which also show the standards used for meeting these objectives in order to achieve the full Semantic Web vision. The stack is still evolving and it is still not clear how the top of the stack can be realized.

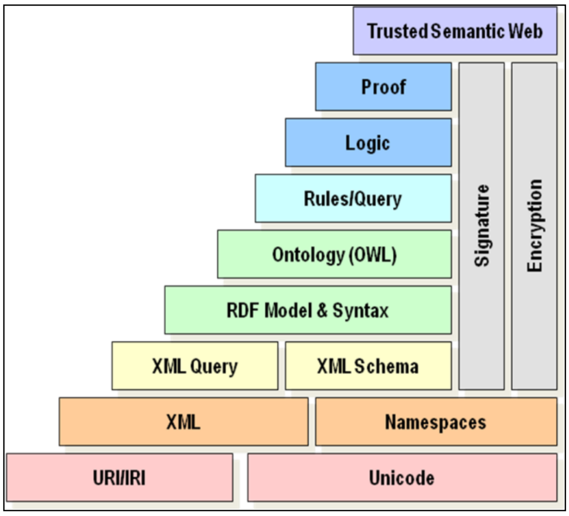


Figure 7: The stack of Semantic Web technologies

A more detailed figure follows which also shows the Linked Open Data, which is described next.

The figure shows that the Linked Data Approach adopts a small fraction of the Semantic Web technologies. For the purpose of this document, we use the term “Semantic Web” to refer to the full suite of W3C standards including XML, RDF, SPARQL, query language, and OWL ontology language and so on.

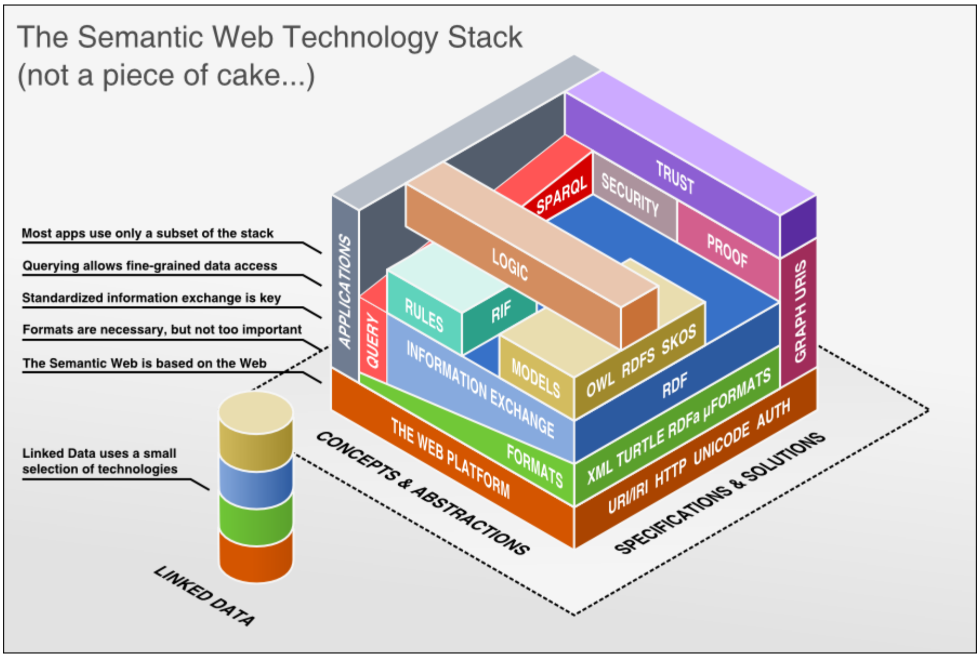


Figure 8: Semantic Web technologies and Linked Data

2.3.1 The Linked open data approach

As shown in Figure 8, the Linked Data approach lies at the core of the Semantic Web stack by adopting a small selection of semantic web technologies. This approach was started as a way of enabling the creation of a **web of data** (a global information space of interlinked datasets, as represented by the Linked Data Cloud in Figure 9), which has the potential to reproduce the network effect that the web of documents had on hypertexts making easier to reuse and combine data on the Web. The rationale behind this approach is that linking data distributed across the Web requires a standard mechanism for describing the existence and meaning of links between the entities named in this data.

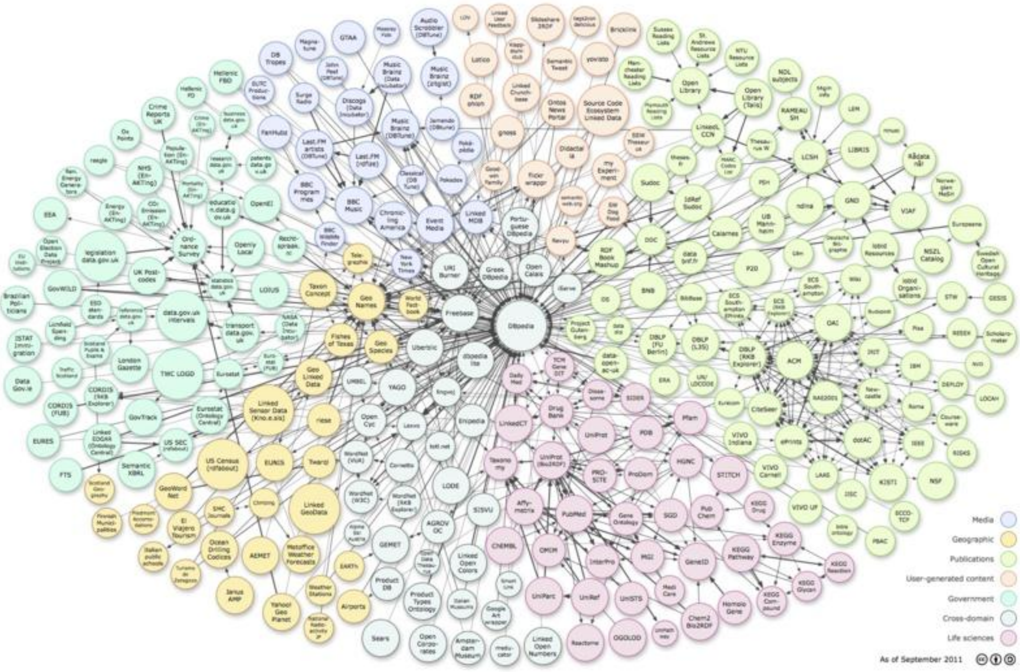


Figure 9: Linked Data Cloud

This idea is summarized by a set of best practices, introduced by Tim Berners-Lee in (Berners-Lee, 2006), for publishing and interlinking structured data on the Web:

1. Use URIs as names of things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs, so that they can discover more things.

The first principle states that URIs, and more specifically HTTP URIs, should be used as a globally unique identification mechanism to name and describe any entity on the Web (not only digital objects but also real world objects and abstract concepts). According to the second principle, the use of HTTP URIs is recommended to enable the resolution of the URIs, over the HTTP protocol, into a description of the identified entities. The key advantage of HTTP URIs is that they can be directly looked up through a pervasive protocol, HTTP, without adding any resolvers or levels of indirection. The third best practice advocates the use of a standard mechanism for publishing structured data on the Web, i.e. the Resource Description Framework (RDF) and a standard language for querying these data, i.e. SPARQL. Finally, the fourth principle, recommends the use of hyperlinks (i.e. RDF links) to connect things and describe relationships between them. RDF links connect distributed data into a single global information space enabling applications to navigate this space and retrieve integrated information about the described entities. The idea is that by representing contents in standardized highly structured formats like RDF, machines can “understand” the meaning of these contents. This would allow the implementation of more sophisticated services, like for services for information searching, which may provide more accurate search results.

In summary, we can state that the Linked Data vision is built upon two simple ideas:

1. To employ the RDF data model to publish structured data on the Web;

2. To use http URIs to describe the properties and relationships between data items across data sources.

Based on this definition, we can conceive the Linked Data principle as an enabling factor for realizing the global web of structured data, which is part of the Semantic Web vision. The key idea of the approach is that the different parts of the resulting Giant Global Graph (GGG) of interconnected data can come from different data sources by allowing the development of intelligent information and content integration services.

According to many, Linked Data seems to offer a valid approach to address many of the interoperability issues faced in digital preservation providing solutions for data search, virtualisation and integration. In the domain of cultural heritage, for example, Semantic Web technologies have been proposed as a promising approach for making multi-format, multi-topical, multi-lingual, multi-cultural and multi-target content mutually interoperable (at syntactic but especially at semantic and organizational levels) so that it can be searched, linked and published in a harmonized way across the boundaries of the datasets and data silos (Hyvönen, 2012).

However, by creating widely-distributed, highly-volatile and largely-integrated streams of data, the Linked Data approach challenges existing paradigms of data sharing and raises the need to re-think provenance and authenticity solutions.

For others, Linked Data has some crucial limitations that make its implementation and use in many contexts very challenging (or even not feasible). A significant example dealing with persistent identifiers of digital resources, as a key of access to digital contents, is described in the next paragraph.

Another raised problem concern the maturity and scalability of Linked Data technologies, which can represent an obstacle for digital preservation that deals with high volumes of data and metadata and has strong persistence requirements. The lack of vocabularies to describe a lot of needed preservation metadata is another mentioned issue by the opponents of the Linked Data approach, even though some solutions are now available such as OAI-ORE to describe the structure of complex objects (i.e. aggregations of Web resources), the PREMIS Data Dictionary for preservation metadata, preservation vocabularies promulgated by the Library of Congress, PRONOM, a public file format registry to support digital preservation services (see Section 4.2 for a detailed description of these solutions). A further problem concerns intellectual property rights and legal issues. Some data in digital preservation institutions like libraries or archives has restricted usage based on local policies, contracts and conditions and rights issues may vary significantly from country to country. Data ownership and rights issues may therefore hinder the release of these data as Linked Open Data.

2.3.2 Coolness vs Persistence: Persistent Identifiers and Linked Data

As explained above, one of the principles to realize the Linked Data vision concerns the use of http URIs to name (i.e. identify) any resource on the Web. Even though this principle was introduced as a way to enabling the creation of a global interlinked information space, recently the practice about HTTP URIs has been proposed as an alternative solution to persistent identifiers for digital objects, authors and other entities relevant for digital preservation and e-Science infrastructures.

This has opened a lively debate between the Linked Data and the Persistent Identifiers (e.g. DOI, URN, ARK, PURL)[5] communities about the implications of adopting the Linked Data practices and tools to extend the value of data (in particular through cross-linking) and cover use cases, which traditional solutions were not designed to address (for example identifying dynamic resources, like datasets and liquid publications).

[5] We refer to D22.1, available at http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D22_1-01-1_9.pdf , for a detailed description of current Persistent Identifiers solutions and approaches.

In order to understand the debate it is important to introduce one important concept of the Linked Data approach, that of **Cool URIs**[6]. A Cool URI is a URI, which does not change. Two things are important about Cool URIs.

1. A naming scheme is necessary to distinguish between resources and their descriptions
2. Cool URIs should be designed to be as stable as possible

Since on the Semantic Web, URIs identify not just Web documents but also real-world entities, like people, organizations and even abstract concepts, a mechanism that establishes what a given URI identifies is mandatory to achieve interoperability between independent information systems. This can be realized through a look-up mechanism that can distinguish between identifiers for Web documents (for example, the person's web page) and identifiers for resources (for example, the identifier for a person). The intended relationship between a resource and its representing documents is depicted in **Figure 10Figure 10: Resolution of Cool URIs**

This can be implemented using two different technical solutions. The first solution is to use a special HTTP status code, "303 See Other", to distinguish non-document resources from regular web documents. The second solution consists of using "hash URIs" for non-document resources.

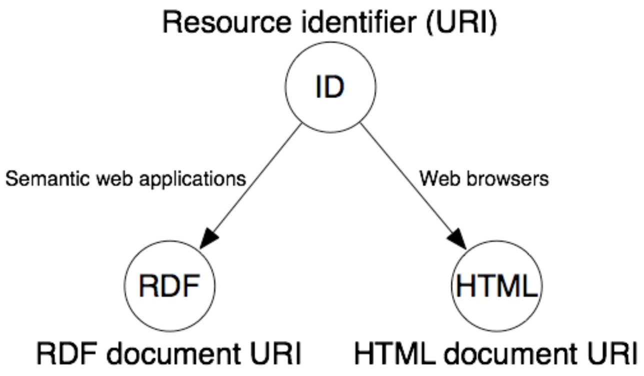


Figure 10: Resolution of Cool URIs

The second aspect about Cool URIs is that they should be as stable as possible. They should be designed in advance and never change following the criteria of simplicity, stability and manageability. Tim Berners-Lee made this point very strongly in his note entitled "Cool URIs don't Change"[7]. The document starts with the statements that "[t]here are no reasons at all in theory for people to change URIs (or stop maintaining documents), but millions of reasons in practice", and the conclusion is: "Keeping URIs so that they will still be around in 2, 20 or 200 or even 2000 years is clearly not as simple as it sounds. However, all over the Web, webmasters are making decisions which will make it really difficult for them in the future. Often, this is because they are using tools whose task is seen as to present the best site in the moment, and no one has evaluated what will happen to the links when things change. The message here is, however, that many, many things can change and your URIs can and should stay the same. They only can if you think about how you design them."

[6] See http://www.w3.org/TR/cooluris/

[7] See http://www.w3.org/Provider/Style/URI

The last point shows some level of awareness in the Linked Data community about the persistence of identifiers. And indeed this approach is sometimes perceived as an easier and faster solution to providing an infrastructure for persistent identifiers. However, from the standpoint of the Persistent Identifiers community, the approach has some weaknesses:

- There is no formal authority, which guarantees the management of Cool URIs (e.g. anyone can assign a Cool URI to anything) and the lifecycle of the identified resources. This means that despite the fact that Cool URIs should not change, the persistence of a HTTP URI as is not guaranteed since no one has the formal responsibility of avoiding changes in URIs (except for the commitment of each web master to put in place solutions which may prevent this from happening). Moreover, Some URI are poorly constructed including components that people may want to change over time (e.g. brand names) and this leads to the proliferation of non-functional links.
- Not only URIs may change but also the identified resources. Resources identified through HTTP URIs may be dynamic, which means that if one resolves the same URI at different times the result can be different. This in general is not a desirable property, because – for example – makes citation very unstable.
- Resources, in particular non-informational ones (like people) are typically identified by different HTTP URIs in each dataset where they are named (this depends on the fact that HTTP URIs encode the domain name as part of the string composing the identifier). The lack of reliable and trustworthy services for mapping these different identifiers onto each other (the current practice is to use point-to-point identity statements between equivalent URIs, but this approach is not very scalable, quite hard to maintain through time) makes the use of HTTP URIs very difficult in the domain of e-Science.

It is important to mention that in the Persistent Identifiers community is slowly emerging a more positive attitude towards the Linked Data vision, which tries to go beyond the vision of Cool URIs as a threat and to identify potential lines of collaboration and reciprocal improvement[8] (Lunghi, Bellini, & Cirinnà, 2012), (Bazzanella, Bortoli, & Bouquet, 2013). In particular, themes which appear to be suitable for such a collaboration are for example the methods for making sure PID's can be referred to as HTTP URI's (including content negotiation), the use of Linked Data vocabularies for naming elements in metadata schemas, the use of owl:SameAs (identity) relations to help identifiers interoperability across PID systems/schema's. In return, there is an expressed will to work with the Linked Data community on defining simple policies/procedures to improve the persistence of HTTP URIs.

---

[8] Among recent initiatives which aims at bridging Persistent Identifiers and Linked Open Data principles is worth to mention the "Den Haag Manifesto", defined during the seminar on Persistent Identifiers organized by Knowledge Exchange on June 14 an d15 2011, as a coordinated effort to identifiers issues across the Persistent Identifier and Linked Data communities. (see: http://www.knowledge-exchange.info/Default.aspx?ID=462)

Another recent event on the same topic was the briefing day titled "Links that last" held in Cambridge on 19/07/2012 to discuss the topics of Persistent Identifiers and Linked Data, discussing practical implications of both approaches for digital preservation.

The international workshop "Interoperability of Persistent Identifier Systems" held on December 13 2012 in Florence paid attention to an interoperability framework for PI systems, see: http://www.rinascimento-digitale.it/workshopPI2012.phtml.

2.4 OVERVIEW OF INTEROPERABILITY INITIATIVES AND PROJECTS FOR DIGITAL PRESERVATION

This section gives an overview of ongoing and past projects and initiatives covering interoperability issues related to digital preservation. To facilitate the reading we organized the descriptions by using a set of categories:

1. **Name**: the name of the initiative/project.
2. **Domain**: indicates a specific area to which the project/initiative belongs.
3. **Timescale**: indicates the duration of the project/ initiative.
4. **Description:** provides information about the project/initiative, its objectives and the issues addressed by it.
5. **Interoperability objectives:** provides a list of interoperability goals addressed by the project/initiative.
6. **Link:** is the URL of a Website where more information and documents can be found.

The projects and initiatives described in this section have been organized around eight macro areas as shown in

Figure 11. The same macro-areas are also listed in **Table 2** where a brief description is provided by keywords to facilitate the navigation of the graph.

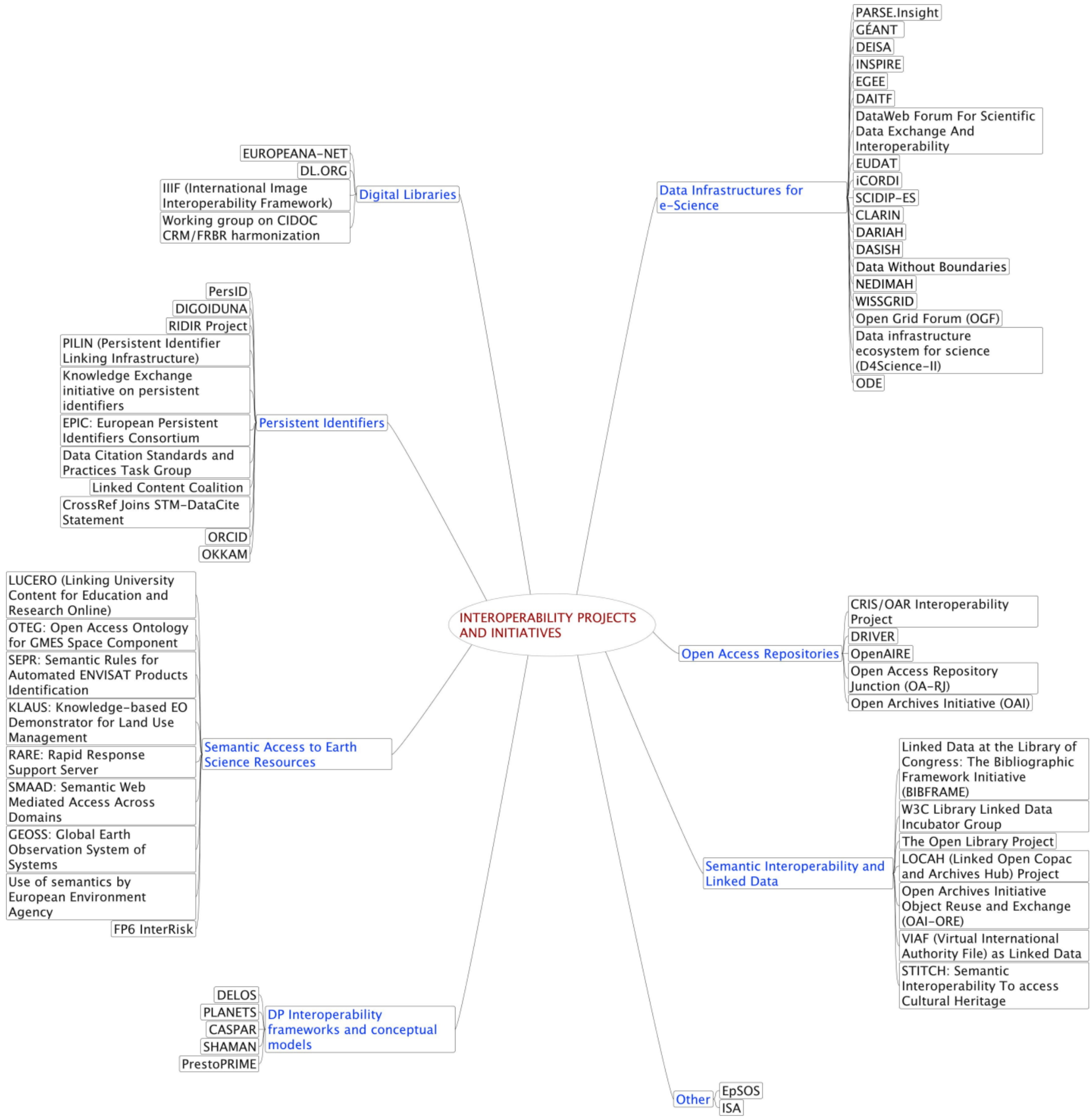


Figure 11: Interoperability Projects and Initiatives

N	DOMAIN	NAME	KEYWORDS
1	Digital Preservation Conceptual Model/ Interoperability Framework	**DELOS**	Digital Library Reference Model
2	Digital Preservation Conceptual Model/ Interoperability Framework	**PLANETS**	PLANETS Interoperability Framework, DP tools and services
3	Digital Preservation Conceptual Model/ Interoperability Framework	**CASPAR**	CASPAR framework, OAIS reference model, virtualization services
4	Digital Preservation Conceptual Model/ Interoperability Framework	**SHAMAN**	SHAMAN DP Open Distributed Resource Management Infrastructure Framework
5	Digital Preservation Conceptual Model/ Interoperability Framework	**PrestoPRIME**	Long-term preservation of digital media objects
6	Data Infrastructure for E-Science: research network	**GÉANT**	e-Science infrastructure for research and education
7	Data Infrastructure for E-Science	**Distributed European Infrastructure for Supercomputing Applications (DEISA)**	Infrastructure for supercomputing applications
8	Data Infrastructure for E-Science: Spatial information	**INSPIRE**	European infrastructure for spatial information
9	Data Infrastructure for E-Science	**Enabling Grids for E-science (EGEE)**	Grid infrastructure for e-Science
10	Data Infrastructure for E-Science	**DAITF: Data Access and Interoperability Task Force**	Data infrastructure for enhance accessibility of data
11	Data Infrastructure for E-Science	**DataWeb Forum For Scientific Data Exchange And Interoperability**	Scientific Data exchange and interoperation
12	Data Infrastructure for E-Science	**EUDAT**	Cross-national and cross-disciplinary collaborative infrastructure
13	Data Infrastructure for E-Science	**iCORDI**	Global data infrastructure
14	Data Infrastructure for E-Science	**SCIDIP-ES**	Data preservation infrastructure for e-

			Science
15	Data Infrastructure for E-Science	**CLARIN**	Research infrastructure for language resources
16	Data Infrastructure for E-Science	**DARIAH**	Research infrastructure of digital humanities
17	Data Infrastructure for E-Science	**DASISH**	Data Service infrastructure for social science and humanities
18	Data Infrastructure for E-Science	**Data Without Boundaries**	Access to official microdata for the European Research Area
19	Data infrastructure for E-Science	**NEDIMAH**	Network for digital methods in the Arts and Humanities
20	Data infrastructure for E-Science	**WISSGRID**	Cross-disciplinary data curation tools for a grid environment
21	Data Infrastructure for E-Science	**Open Grid Forum (OGF)**	Grid computing standardization and community
22	Data Infrastructure for E-Science	**Data infrastructure ecosystem for science (D4Science-II)**	Integrated infrastructure for e-Science
23	Data Infrastructure for E-Science	**Opportunities for Data Exchange**	Interoperable data sharing, re-use and preservation layer for e-Science
24	Digital Libraries	**EUROPEANA-NET**	European Digital Library
25	Digital Libraries	**DL.ORG**	Interoperability challenges in digital libraries
26	Digital Libraries	**IIIF**	Interoperability framework for image delivery
27	Digital Libraries/Museums	**Working Group on FRBR/CRM Harmonization**	Object-oriented FRBR (FRBRoo)
28	Open Repositories	**CRIS/OAR Interoperability Project**	Metadata exchange format
29	Open Repositories	**DRIVER and DRIVER II**	Infrastructure of digital repositories
30	Open Repositories	**OpenAIRE**	Infrastructure for Open Access publishing of research output
31	Open Repositories	**Open Access Repository**	Deposit broker tool

		Junction (OA-RJ)	
32	Open Repositories	**Open Archives Initiative (OAI)**	Interoperability standards for content access and dissemination
33	Open Repositories	**Open Archives Initiative Object Reuse and Exchange (OAI-ORE)**	Standards for the description and exchange of Web resource aggregations
34	Persistent Identifiers	**PersID**	Persistent identifier infrastructure for digital publications and digital resources.
35	Persistent Identifiers	**DIGOIDUNA**	Digital Objects and Author Identifiers study
36	Persistent Identifiers	**RIDIR Project**	Interoperability between digital repositories through PIDs
37	Persistent Identifiers	**PILIN (Persistent Identifier Linking Infrastructure)**	Global PIDs infrastructure
38	Persistent Identifiers	**Knowledge Exchange initiative on persistent identifiers**	International Infrastructure based on URN/NBN
39	Persistent Identifiers	**EPIC: European Persistent Identifiers Consortium**	PIDs service based on Handle
40	Persistent Identifiers	**Data Citation Standards and Practices Task Group**	Data citation coordination, guidelines and best practices
41	Persistent Identifiers	**Linked Content Coalition**	Infrastructure for digital rights management
42	Persistent Identifiers	**CrossRef Joins STM-DataCite Statement**	Interoperability between CrossRef and DataCite services
43	Persistent Identifiers	**ORCID**	Unique IDs for researchers and contributors
44	Persistent Identifiers	**OKKAM**	Global Unique IDs for digital and non-digital entities; Entity Name System (ENS)
45	Semantic Interoperability and Linked Data	**Linked Data at the Library of Congress: The Bibliographic Framework Initiative**	Convert MARC21 to Linked data

		(BIBFRAME)	
46	Semantic Interoperability and Linked Data	**W3C Library Linked Data Incubator Group**	Standardization in library community
47	Semantic Interoperability and Linked Data	**The Open Library Project**	Linked Open Data for bibliographic items
48	Semantic Interoperability and Linked Data	**LOCAH (Linked Open Copac and Archives Hub) Project**	Aggregation and merging of Hub data
49	Semantic Interoperability and Linked Data	**VIAF (Virtual International Authority File) as Linked Data**	Global authority service
50	Semantic Interoperability and Linked Data	**STITCH: Semantic Interoperability To access Cultural Heritage**	Semantic interoperability for metadata vocabularies (STITCH is a CATCH project (see below))
51	Semantic Interoperability and Linked Data	**CATCH program (Continuous Access to Cultural Heritage)**	Generic methods and techniques cutting across the areas of the humanities and computer science, aiming to facilitate an interaction with cultural heritage institutions.
52	Semantic interoperability and Linked Data	**CEDAR (Linked Open Census Data)**	Semantic Data Web of Historical Information (Census data)
53	Semantic Interoperability and Linked Data	**LUCERO (Linking** University Content for Education and Research Online)	Open University datasets interoperability
54	Semantic Access to Earth Sciences resources	**OTE: Ontology and Terminology for Earth Observation**	Earth Observation (EO) ontology mapping
55	Semantic Access to Earth Sciences resources	**OTEG: Open Access Ontology for GMES Space Component**	Extending OTE results, overall semantic architecture in the EO domain
56	Semantic Access to Earth Sciences resources	**SEPR: Semantic Rules for Automated ENVISAT Products Identification**	Ways of mapping between application domains and EO resources
57	Semantic Access to Earth Sciences resources	**KLAUS: Knowledge-based EO Demonstrator for Land Use Management**	Ontology search

58	Semantic Access to Earth Sciences resources	**RARE: Rapid Response Support Server**	Distributed software system accessible through a Web-based user interface that allows searching for EO-related resources
59	Semantic Access to Earth Sciences resources	**SMAAD: Semantic Web Mediated Access Across Domains**	Semantic discovery of EO resources
60	Semantic Access to Earth Sciences resources	**GEOSS: Global Earth Observation System of Systems**	Exchange data (and metadata) between heterogeneous sources
61	Semantic Access to Earth Sciences resources	**Use of semantics by European Environment Agency**	Shared Environmental Information space
62	Semantic Access to Earth Sciences resources	**FP6 InterRisk**	Access to information for risk management
63	Other (eHealth)	**EpSOS**	eHealth
64	Other (eGovernment)	**ISA**	Interoperability between European public administrations

Table 2: List of Projects and Initiatives described in D25.1

2.4.1 Digital Preservation Shared Conceptual Models and Interoperability Frameworks

The early research projects in the field of digital preservation focused on the definition of basic concepts and shared conceptual models as fundamental ways to enable interoperability of the various content holders - where interoperability is intended as the ability to access and reuse information across collections in diverse media and languages, administered independently - and rising awareness about the theoretic basis for the key preservation concepts and entities. The projects were mainly driven by the digital library and archive community. DELOS is one of the project of this early phase which contributed to the definition of a digital library reference model representing a first step toward the development of an European Digital Library.

In the next stage, interoperability has been addressed by developing solutions to integrate digital preservation modules into framework architectures to enable the interoperation with other systems. Examples of this kind of architectures are the PLANETS Interoperability Framework, the CASPAR Integrated Framework and the integrated preservation framework using grid-technologies of SHAMAN. In this section we give an overview about the main digital preservation projects, which addressed interoperability issues by defining shared conceptual models or developing interoperability framework architectures (for an overview about EU funded research projects on digital preservation see (Strodl, Petrov, & Rauber, 2011)).

Name	**DELOS**
Domain	Network of Excellence on Digital Preservation (DP)

Timescale	2004 - 2008
Description	DELOS is a Network of Excellence partially funded by the EC with the aim to coordinate and integrate the research efforts in the domain of digital libraries. The final goal of this network of excellence is to develop the next generation of digital libraries technologies, based on comprehensive models and frameworks for the life-cycle of digital library objects.
Interoperability objectives	An important research challenge of DELOS has been the interoperability of the various content holders to allow access to information across collections managed by independent systems. The Digital Library Reference Model is the main DELOS contribution to this domain. The aim of the model is to provide the formal and conceptual basis for the design and implementation of a Digital Library Management System, describing the characteristics of the main functionalities and actors involved.
Link	http://www.delos.info/

Name	**PLANETS (**Preservation and Long-term Access through Networked Services)
Domain	Digital Preservation (DP)
Timescale	2006 -2010
Description	PLANETS is a four-year project co-funded by the European Union with the primary goal of building services and tools for the long-term access to digital cultural and scientific assets.
Interoperability objectives	• Developing an integrated architecture for digital preservation actions: the Planets Interoperability Framework (more details can be found in Section 4.2). The approach is to provide a framework and services that can be integrated with existing systems (e.g. local archiving storage systems). • Interoperability benefits of the framework: 1) It enforces semantic interoperability (digital object model) between various services of a preservation workflow and provides a number of commonly required software components by re-using existing preservation patterns. 2) By providing common components, the IF can also help to assure that various applications remain interoperable. 3) By enforcing Web Service standards, the IF can support access to remote and distributed third-party characterization and migration services.
Link	http://www.planets-project.eu/

Name	**CASPAR** (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval)
Domain	Digital Preservation (DP)
Timescale	2006 -2010
Description	CASPAR is an Integrated project co-funded by the European Union with the ambitious challenge of building up a common preservation framework for heterogeneous data based on the implementation, extension and validation of the OAIS reference model. The CASPAR framework has been developed on the following basic interoperability principles: 1) compliance with the main standard of reference in DP, the OAIS Reference Model; 2) technology neutrality; 3) loosely-coupled architecture; 4) domain Independence.

Interoperability objectives	• to establish the foundation methodology and a reference framework, based on the OAIS reference model, for covering all preservation aspects. • Design virtualisation services supporting long-term digital resource preservation, despite changes in the underlying computing (hardware and software) and storage systems, and the Designated Communities. • to validate the CASPAR approach to digital resource preservation across different user communities. • Actively contribute to the relevant standardisation activities in areas addressed by CASPAR. • Raise awareness about the critical importance of digital preservation among the relevant user-communities and facilitate the emergence of a more diverse offer of systems and services for preservation of digital resources.
Link	http://www.casparpreserves.eu/

Name	**SHAMAN** (Sustaining Heritage Access through Multivalent Archiving)
Domain	Digital Preservation (DP)
Timescale	2007 -2011
Description	SHAMAN is large integrated project co-financed by the EU that aims to develop and test a next generation digital preservation framework including tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives.
Interoperability objectives	• To develop a next generation distributed digital preservation management framework, with the respective tools for the analysis, ingestion, management, access and re-use of information objects and data across libraries and archives; The infrastructure enables grid-based resource integration, extending the OAIS model and taking advantage of the latest state of the art in virtualization and distribution technologies from the field of GRID computing, Federated Digital Libraries and Persistent Archives. • To enable standard services to be executed across independently managed computational and storage resources. • To demonstrate the feasibility to automate preservation management policies which are needed in an interoperability scenario between institutions. • To support the managing of future requirements by securing interoperability with future environments and maintaining essential properties of the preserved content.
Link	http://shaman-ip.eu/start

Name	**PrestoPRIME**
Domain	Digital Preservation (DP)
Timescale	2009 - 2013
Description	PrestoPRIME is an European project aiming at developing practical solutions for the long-term preservation of digital media objects.
Interoperability	• developing (and implementing a prototype of) a preservation framework

objectives	including risk management and content quality and corruption control measures, capable of supporting audiovisual signal migration and multivalent preservation methods using federated services for distributing and storing content. • Establishing metadata interoperability between audiovisual archives, cultural heritage institutions, the Semantic Web and content portals, with services for metadata conversion and deployment. • Developing a rights management system and audiovisual fingerprint registry to track and manage content at all stages of its lifecycle, in all contexts of use. • To establish a European networked Competence Centre to support community interoperability gathering the knowledge created by PrestoPRIME and delivering advanced digital preservation advice and services in conjunction with the European Digital Library Foundation and other projects.
Link	http://www.prestoprime.org/project/index.en.html

2.4.2 Data Infrastructure for e-Science

E-science infrastructures are widely recognized as one of the main pillars of research and innovation in Europe[9] providing new ways of collaboration and resource sharing in key areas of science. However a closer look reveals that this wide variety of heterogeneous infrastructures raises several interoperability and preservation challenges both within and across system boundaries. Therefore, the creation of an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-Science infrastructures has been recognized as a fundamental pre-requisite for the realization of their innovation potential (ODE project). The PARSE.Insight project[10], for example, showed a substantial demand for a science data infrastructure, which is consistent across nations, continents and disciplines. The project, described in more details in this section, identified a number of specific components, both technical and non-technical, which should be needed to supplement existing and already planned infrastructures for science data in order to bridging the gaps between islands of functionalities and allow interoperability across national and disciplinary boundaries.

This section reviews existing initiatives that aim to promote interoperability in specific e-science domains through the implementation of such infrastructures and describe also some relevant initiatives committed to promote and develop reference models and architectures to enable infrastructure interoperability across systems.

Name	**PARSE.Insight**
Domain	Digital preservation in e-Science infrastructures
Timescale	2008-2010
Description	PARSE.Insight was a two-year project co-funded by the EU under the Seventh Framework Programme concerned with preservation of digital information in science. In particular, the project focused on the parts of e-Science infrastructure needed to support persistence and understandability of the digital assets of EU

9 http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what

10 see D2.1, Draft Road Map, available at http://www.parse-insight.eu/publications.php

	research. Technical and non-technical components which would be needed to supplement existing and already planned infrastructures for science data are described in the PARSE.Insight roadmap. These components are intended to play an important unifying role in enabling interoperability in science data.
Interoperability objectives	• To develop a shared vision and framework for a sustainable interoperable organizational infrastructure for permanent access to scientific information. • a roadmap for the support e-infrastructure for long-term accessibility and usability of scientific and other digital information in Europe; • to share and capitalise on best practices as well as understanding the impact that e-Science is having on the research communities that it is serving; • an international process for evaluating the sustainability and trustworthiness of digital repositories.
Link	http://www.parse-insight.eu/project.php

Name	**GÉANT**
Domain	Research network and e-Science
Timescale	2000-2004 (GN1) 2004-2009 (GN2) 2009-2013 (GN3)
Description	The project started as collaboration between 26 National Research and Education Networks (NRENs) across Europe, the European Commission, and DANTE (Delivery of Advanced Network Technology to Europe). The GÉANT project's purpose has been to improve the previous TEN-155 pan-European research network by creating a new pan-European infrastructure for research and education community– the GÉANT network. This network became fully operational on 1 December 2001. With the infrastructure in place, the current GÉANT project is focused on developing tools and services in collaboration with Europe's NRENs, aimed at better-serving the evolving needs of the research and education community into the long term.
Interoperability objectives	Interoperability objectives can be divided in 3 kinds of activities within the project: 1. Research activities: critical analysis of future network technologies in order to promote interoperable services and solution to secure persistent access to network resources. 2. Services activities: building the pan-European network and multi-domain services 3. Network activities: enabling community interoperability by seeking to understand needs and requirements of the user community, facilitating the dissemination of the network, services and other activities at European and worldwide level.
Link	http://www.geant.net/

Name	**Distributed European Infrastructure for Supercomputing Applications (DEISA)**
Domain	e-Science infrastructures
Timescale	2002-2008 (DEISA initiative, DEISA I and II)

(start date)	
Description	DEISA is a European Union supercomputing project, which involved a consortium of eleven national supercomputing centres. The project aim was to develop a distributed high performance computing infrastructure in Europe.
Interoperability objectives	• Deployment of a cooperative European HPC ecosystem built on top of national service.
Link	http://www.deisa.eu/

Name	**INSPIRE**
Domain	Infrastructure for Spatial Information
Timescale	2007- 2019
Description	INSPIRE (Infrastructure for Spatial Information in the European Community) is a EU initiative aiming at establishing an infrastructure for spatial information in Europe that will help to make spatial or geographical information more accessible and interoperable for a wide range of purposes. Since 2007 its growing community of geographic information specialists have come up with a conceptual model for developing interoperability specification in spatial data infrastructure. According to INSPIRE, interoperability is defined as “the possibility for spatial datasets to be combined, and for services to interact, without repetitive manual intervention, in such a way that the result is coherent and the added value of the datasets and services is enhanced”. In order to ensure interoperability of special data of Member States, the INSPIRE directive requires that common implementing rules are adopted in the following areas: 1) Metadata 2) Data specification 3) Network services 4) Data and Service Sharing 5) Monitoring and Reporting
Interoperability objectives	• To share environmental spatial information among public sector organisations • To combine seamless spatial information from different sources across Europe and share it with many users and applications • To share information collected at one level/scale with all levels/scales • To facilitate public access to spatial information across Europe • Policy-making across boundaries
Link	http://inspire.jrc.ec.europa.eu/

Name	**Enabling Grids for E-science (EGEE)**
Domain	Grid infrastructure for e-science
Timescale	2004-2010
Description	The aim of the project was to develop a collaborative production grid infrastructure for e-science.
Interoperability objectives	• The creation of the largest collaborative production grid infrastructure in the world for e-science, demonstrating that such a production infrastructure

	can be used by a wide range of research disciplines, and producing scientific results in these disciplines which would otherwise not have been possible.
Link	http://www.eu-egee.org/

Name	**DAITF: Data Access and Interoperability Task Force**
Domain	Data infrastructures for e-science
Timescale (start date)	October 2012
Description	DAITF is an international group, which involves researchers, operators of elements of data infrastructure, and policy makers, devoted to promoting incremental improvements in data infrastructures to enhance the accessibility and interoperability of data at a global scale. The main objective of this group is to ensure the development of data infrastructures, which allow maximum access and interoperability across system boundaries.
Interoperability objectives	• Promoting incremental improvements in data infrastructures to enhance the accessibility and interoperability of data at a global scale.
Link	http://www.daitf.org/

Name	**DataWeb Forum For Scientific Data Exchange And Interoperability**
Domain	Scientific data exchange
Timescale (Start date)	2012
Description	The DataWeb forum is a non-profit voluntary organization the foundation of which has been recently proposed in a concept paper[11] by Chris Greer from the National Institute of Standards and Technology and Alan Blatecky of the National Science Foundation. The idea behind the DataWeb Forum is to establish an organization to facilitate the exchange and interoperation of scientific data across disciplines and national boundaries by creating an open, seamless, self-regulatory global digital data infrastructure that is the foundation for discovery and progress. The DWF, which is envisioned not as a standard body but as an open voluntary organization, will pursue its mission for harmonizing standards, policies and technologies, and other implementation elements primarily through the development of action plans and associated working groups.
Interoperability objectives	• To facilitate the exchange and interoperability of data across disciplines and national boundaries by producing high quality, relevant technical documents that influence the way people store, use, and manage data.
Link	http://www.cni.org/news/dataweb-forum-for-scientific-data-exchange-and-interoperability/

Name	**EUDAT**

[11] http://www.cni.org/wp-content/uploads/2012/06/DataWebForum_Concept_Paper.pdf

Domain	**Data infrastructures for e-science**
Timescale	October 2011 – September 2014
Description	The pan-European data initiative EUDAT brings together a unique consortium of 25 European partners – including research communities, national data and high performance computing centres, technology providers, and funding agencies. EUDAT aims at building a sustainable cross-disciplinary and cross-national Collaborative Data Infrastructure (CDI) in which centres offering community-specific support services to their users could rely on a set of common data services shared between different research communities. The main focus of EUDAT will be on building this common layer of generic cross-disciplinary data services.
Interoperability objectives	• Interoperability between data infrastructures
Link	http://www.eudat.eu/

Name	**iCORDI**
Domain	Data infrastructures for e-science
Timescale	September 2012 – August 2014
Description	iCORDI is an EU project aiming to be the premium global forum driving convergence between emerging global data infrastructures. Its prime objective is to establish a coordination platform between Europe, the USA and beyond to discuss and improve the interoperability of today's and tomorrow's scientific data infrastructures of both continents and to extend this to the global levels. iCORDI supports the creation of the Research Data Alliance (RDA). The RDA is an emerging organization that will facilitate specific, short-term efforts that accelerate the sharing and exchange of research data. The purpose of RDA is to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability. This will be achieved through the development and adoption of infrastructure, policy, practice, standards, and other deliverables.
Interoperability objectives	• Foster discussion between relevant stakeholders in the EU and US on concrete topics related to interoperability of the data infrastructures and solutions based on a top-down approach. • Overcome identified challenges and turn ideas on convergence into concrete specifications that can be immediately implemented in both continents by bringing data practitioners together in a bottom-up process. • Demonstrate what can realistically work and pinpoint remaining barriers that have to be tackled to achieve full interoperability.
Link	http://www.icordi.eu

Name	**SCIDIP-ES**
Domain	e-science infrastructure (Earth Science)
Timescale	October 2011 – September 2014
Description	The SCIence Data Infrastructure for Preservation – Earth Science (SCIDIP-ES) is a EU project that aims to develop a data preservation infrastructure for e-science by

	delivering generic services for science data preservation as part of the data infrastructure for e-science. Building on the experience of the ESA Earth Observation Long Term Data Preservation (LTDP) programme the initiative aims to favour the set-up of a European Framework for the long term preservation of Earth Science (ES) data through the definition of common preservation policies, the harmonization of metadata and semantics and the deployment of the generic infrastructure services in the ES domain.
Interoperability objectives	• Harmonization of metadata and semantics. • Harmonizing data preservation policies, approaches and tools in the Earth Science Domain. • Definition of a European Earth Science Infrastructure architecture and Governance Model
Link	http://www.scidip-es.eu/science-data-infrastructure-for-preservation-with-focus-on-earth-science/

Name	**CLARIN**
Domain	Research infrastructure of Language resources
Timescale (start date)	2009-2015
Description	The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable for the whole European Humanities (and Social Sciences) community.
Interoperability objectives	• To establish an integrated and interoperable research infrastructure of language resources and its technology.
Link	http://www.clarin.eu

Name	**DARIAH**
Domain	Research infrastructure for the Arts and Humanities
Timescale (start date)	2012
Description	The grand vision for DARIAH is to facilitate long-term access to, and use of, all European Arts and Humanities digital research data. The DARIAH infrastructure will be a connected network of people, information, tools, and methodologies for investigating, exploring and supporting work across the broad spectrum of the digital humanities. The core strategy of DARIAH is to bring together national, regional, and local endeavours to form a cooperative infrastructure where complementarities and new challenges are clearly identified and acted upon.
Interoperability objectives	• To establish an integrated and interoperable research infrastructure for the Arts and Humanities
Link	http://www.dariah.eu

Name	**DASISH**
Domain	Data Service Infrastructure for the Social Sciences and Humanities
Timescale	2012-2014

(start date)	
Description	DASISH project brings together all 5 research infrastructure initiatives in the social sciences and humanities (SSH) represented each by some centers: CLARIN – Common Language Resources and Technology Infrastructure DARIAH – Digital Research Infrastructure for the Arts and Humanities CESSDA – Council of European Social Science Data Archives ESS – European Social Survey SHARE – Survey of Health, Ageing and Retirement in Europe The goal is to determine areas of possible synergies in the infrastructure development and to work on a few concrete joint activities. The rationale behind this idea is that: • Double developments should be prevented • Initiatives should mutually benefit from the advanced work of the others • To establish joint integrated domains where this makes sense for the SSH users
Interoperability objectives	• Joint activities will be along the following dimensions: understanding the different architectural solutions, assessing and improving data and metadata quality, setting up a tools and services forum, improve the quality of survey data, locate and improve data preservation and curation services, develop a joint shared data access and enrichment framework (AAI, PIDs, joint Metadata, Workflow implementations, joint annotation framework), jointly work on legal and ethical aspects, carry out much training and education work, work on disseminating the results.
Link	http://www.dariah.eu

Name	**Data Without Boundaries (DwB)**
Domain	Research infrastructure for the Arts and Humanities
Timescale (start date)	2012-2014
Description	The DwB project supports equal and easy access to official (statistical) microdata for the European Research Area, within a structured framework where responsibilities and liability are equally shared.
Interoperability objectives	• DwB contributes to the creating of an integrated model where the best solutions for access to official statistical microdata are available, irrespective of national boundaries, yet flexible enough to fit national arrangements. • DwB enhances researchers' transnational access to official micro data through coordination of existing infrastructures, Council of European Social Science Data Archives (CESSDA) and the European Statistical System (ESS).
Link	http://www.dwbproject.org/

Name	**NEDIMAH**
Domain	Research Network for Digital Humanities

Timescale (start date)	2011-2015
Description	NeDiMAH (Network for Digital Methods in the Arts and Humanities) is carrying out a series of activities and networking events that will allow the examination of the practice of, and evidence for, digital research in the arts and humanities across Europe. It will build collaborations and networking between the community of European scholars active in this area, as well as those engaged with creating and curating scholarly and cultural heritage digital collections. NeDiMAH activities and research will contribute to the classification and expression of digital arts and humanities via three key outputs: • A map visualising the use of digital research across Europe • An ontology of digital research methods • A collaborative, interactive online forum for the European community of practitioners active in this area
Interoperability objectives	Sustainability of digital humanities services, tools and services is an important aspect of the project. Thus, interoperability is important to improve the sustainability.
Link	http://www.nedimah.eu/

Name	**WissGrid project**
Domain	Data curation tool for grid infrastructures
Timescale (start date)	2009 – 2012
Description	The project aims to provide cross-disciplinary data curation tools for a grid environment by adapting repository concepts and technologies to the existing D-Grid e-Infrastructure.
Interoperability objectives	• to establish long-term organisational and technical D-Grid structures for the academic world • to promote sustainability of scientific data management, its long-term curation and cross-disciplinary re-use • WissGrid aims to support the communities in establishing their own curation strategies and systems, and supports convergence and exchange of experiences between them.
Link	http://www.wissgrid.de/

Name	**Open Grid Forum (OGF)**
Domain	Distributed (Grid) Computing
Timescale (start date)	2006
Description	OGF is an open community which has 2 main objectives: 1) driving the rapid evolution of distributed computing by leading the standardization effort for grid computing, 2) build the grid community, by exploring trends, promoting good practices and consolidate the best practices into standards.
Interoperability objectives	• developing open standards for Grid software interoperability

Link	http://www.ogf.org/

Name	**Data infrastructure ecosystem for science (D4Science-II)**
Domain	e-infrastructures
Timescale	2009-10-01 2011-09-30
Description	D4Science-II is an European e-infrastructure project co-funded by the European Commission's Seventh Framework Programme for Research and Technological Development. The project aims to develop technological solutions to enable interoperability between diverse data e-infrastructures in order to create an integrated e-infrastructure ecosystem for communities from different disciplines. A prototype of this ecosystem for several scientific e-Infrastructures established in areas such as biodiversity, fishery resources management and high energy physics is one of the main goals of the project.
Interoperability objectives	• to promote interoperability across e-infrastructures by empowering large user communities.
Link	http://www.d4science.eu/about

Name	**Opportunities for Data Exchange (ODE)**
Domain	e-Infrastructures
Timescale	2010 2012
Description	ODE is a project proposed by APA and four of its members CERN, Finnish Computer Centre for Science, Helmholtz Association and UK Science and Technology Funding Council. The aim of the project is to identify, collate, interpret and deliver evidence of emerging best practices in sharing, re-using, preserving and citing e-Science data in order to contribute to the creation of an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-Infrastructures.
Interoperability objectives	• Enable operators, funders, designers and users of national and pan-European e-Infrastructures to compare their vision and explore shared opportunities • Demonstrate and improve understanding of best practices in the design of e-Infrastructures leading to more coherent national policies • Document success stories in data sharing, visionary policies to enable data re-use, and the needs and opportunities for interoperability of data layers to fully enable e-Science
Link	http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/

2.4.3 Digital libraries

Interoperability is a central issue in the digital library space where independent systems are demanded to support a broad variety of interdisciplinary activities, data sharing and re-use and address the data deluge of the modern digital era. Considerable advances in digital library technologies and services have been made to address the need for Global Digital Library Infrastructures resulting from the federation of regional, national and international interoperable DL systems and repositories. However making independent digital libraries interoperable is a complex issue that needs to take into considerations many aspects such as the digital library architecture, functionalities, interfaces, policies, rights restrictions, responsibilities and user perspectives and requirements.

In this section, we describe some of the most relevant initiatives to address the interoperability challenges in the domain of digital library. Some of these initiatives focused on the development of a common conceptual framework for enabling interoperability between digital libraries (e.g. DL.ORG) or for exchanging specific types of content (e.g. IIIF), others addressed the issue of creating a unique point of entry to distributed content and heterogeneous resources (e.g. EUROPEANA, EUROPEANA GROUP).

Name	**EUROPEANA-NET (European Digital Library Network)**
Domain	**Digital Library**
Timescale	1 July 2007 – 31 March 2009
Description	EuropeanaNet aimed to tackle the fragmented cultural heritage map of Europe by creating consensus across the four domains of Archives, Museums, Libraries and Audio-visual Archives to be able to build Europeana, the European digital library, museum and archive. The Thematic Network encompassed representatives from these four cultural domains across Europe. During its 21-month lifespan, the project addressed the human, political, semantic, technical and organisational interoperability issues that arise in providing a multi-lingual, single portal for searching within and across Europe's distributed and varied cultural resources.
Interoperability objectives	• Creating a broad community from the four cultural domains reported above to make content available and interoperable, in support of European Digital Library. • A fully working prototype, with interoperable multilingual access to over 4 million digital items.
Link	http://www.europeana.eu

Name	**EUROPEANA GROUP**
Domain	**Digital Library**
Timescale	
Description	They are a number of projects, run by different cultural heritage institutions, that are contributing technology solutions and aggregated content to Europeana.
Interoperability objectives	The majority of the Europeana group projects provide aggregation services to bring content from different sources to Europeana (APEnet, ATHENA, CARARE, Europeana Connect, Europeana Libraries, thinkMOTION, OpenUp, MIMO, Europena Library, Europeana Film Gateway, Europeana Local, Europeana Travel). Some projects work on raising awareness and sharing common policies, tools and standards (ATHENA, Europena Local).
Link	http://pro.europeana.eu/web/europeana-pro

Name	DL.ORG Digital Library Interoperability, Best Practices & Modeling Foundations
Domain	Digital Library
Timescale (start date)	2008

Description	DL.org is a 2-year Coordination Action project, which started December 2008, funded by the Commission of the European Union (EC) under the 7th Framework Programme ICT Thematic Area "Digital Libraries & Technology-Enhanced Learning. DL.org has investigated from many perspectives the interoperability challenges in the domain of digital libraries. The two main outputs of this effort are:
	• The Digital Library Technology & Methodology Cookbook[12] which provides a framework for addressing interoperability coupled with patterns and solutions.
	• The DL-org Digital Library Reference Model[13] including:
	  ○ The Digital Library Manifesto
	  ○ The Digital Library Reference Model- in a Nutshell
	  ○ The Digital Library Reference Model Concepts and Relations
	  ○ The Digital Library Reference Model Conformance Check List

Interoperability objectives	• DL.org investigates interoperability in the context of digital library from an all-encompassing perspective including content, functionalities, policy, quality, user and architecture. Six working groups have been constituted to work on these interoperability issues. • A liaison group of high-profile members of organizations and coalition of DL.org and interoperability stakeholders has been constituted with the aim of evaluating the output of the DL.org working groups, enlarging participation and disseminating the outputs. This effort is fundamental to foster the interoperability at organizational and community levels.
Link	http://www.dlorg.eu/

Name	IIIF International Image Interoperability Framework
Domain	Digital Library
Timescale (start date)	2011
Description	It is a one-year collaborative effort committed by British Library, Stanford University, Bodleian Libraries, Oxford University, La Bibliothèque nationale de France, Nasjonalbiblioteket (National Library of Norway), Los Alamos National Library and Cornell University, aiming at producing an interoperable framework for image delivery.
Interoperability objectives	• The IIIF including shared technology, common API's, rich user interfaces, is designed to give scholars a level of uniform and rich access to image based resources, regardless of the holding institution(s).

[12] The DG.org Cookbook is available at http://www.dlorg.eu/index.php/outcomes/dl-org-cookbook

[13] The DG.org reference model is available at http://www.dlorg.eu/index.php/outcomes/reference-model

	• Metadata interoperability: the IIIF group is working on a metadata specification (mainly structural metadata) to drive common image viewer requirements.
Link	http://lib.stanford.edu/iiif

Name	**International Working Group on FRBR/CIDOC CRM Harmonization**
Domain	Digital Library and Museum
Timescale (start date)	2003
Description	It is a special interest group which that brings together representatives from library and museum communities with the common goal of harmonizing the FRBR model and the CIDOC CRM model.
Interoperability objectives	• Expressing the IFLA FRBR model with the concepts, tools, mechanisms, and notation conventions provided by the CIDOC CRM. • Aligning (possibly even merging) the two object-oriented models with the aim to contribute to the solution of the problem of semantic interoperability between the documentation structures used for library and museum information.
Link	http://www.ifla.org/node/928 http://www.cidoc-crm.org/frbr_inro.html

2.4.4 Open repositories

In the recent years, Open Access repositories and their associated services have become an increasingly important component of e-Science Infrastructures. It has been widely recognized that the real potential of open access repositories for e-Science infrastructures lies on the creation of a network of interconnected repositories providing unified access to distributed scientific resources and scholarly content. The creation of this decentralized infrastructures and the development of added-value services on top of it are entirely reliant on interoperability. “Interoperability is the technical glue that makes it possible to virtually connect repositories to each other and to other systems and transfer information, metadata, and digital objects between each other.”[14]

This section describes projects and initiatives focused on interoperability challenges in the domain of Open Access repositories. It is not intended to be an exhaustive description of all interoperability initiatives in this domain, but rather it provides information about some interoperability initiatives, which addressed interoperability challenges relevant also for digital preservation purposes. Concrete solutions and services (AuthorClaim, OAI-PMH, SWORD) which have been developed as result of these and other initiatives will be described in Section 4.2.

We restricted the focus on initiatives addressing four major issues:

1) Metadata harvesting and exchange
2) Infrastructures for digital repositories
3) Repository deposit

[14] *The Current State of Open Access Repository Interoperability (2012)* available at http://www.coar-repositories.org/files/COAR-Current-State-of-Open-Access-Repository-Interoperability-26-10-2012.pdf

Name	**CRIS/OAR Interoperability Project**
Domain	**Open Access Repositories**
Timescale	2009-2010
Description	The project aims to increase the interoperability between Current Research Information Systems (CRIS) and Open Access Repository (OAR) systems by defining and proposing a metadata exchange format for publication information with an associated common vocabulary.
Interoperability objectives	• Making metadata input more efficient • Avoiding or reducing duplicated inputs on the two platforms • Increasing metadata quality, reliability and reusability • Increasing quality level of services based on these metadata • Reducing costs of metadata handling and exchange
Link	https://infoshare.dtv.dk/twiki/bin/view/KeCrisOar/WebHome

Name	**DRIVER and DRIVER II**
Domain	Digital repositories
Timescale	2007- 2009
Description	DRIVER is a multi-phase effort, which aims at establishing a pan-European infrastructure of digital repositories as an integral part (i.e. the layer which provides the content) of e-infrastructure for research and education in Europe. The general principle behind this initiative is to link users to knowledge. To this aim the plan of DRIVER is to create the basis for a distributed network of repositories across Europe to facilitate information access and search by users. The second phase of the project named DRIVER II aims at improve the results of DRIVER, for example by expanding the geographical coverage of digital repositories included in the network, including different kinds of resources as non-textual or non-publication resources, enabling the construction of enhanced publications.

Interoperability objectives	• an interoperable, trusted, long-term repository infrastructure hub in a global repository network for:
	  o any type of document
	  o of any type of format
	  o involving all European countries
	  o covering all academic disciplines

Link	http://www.driver-repository.eu/

Name	**OpenAIRE**
Domain	Open Access infrastructure for the Arts and Humanities
Timescale (start date)	2009-2012
Description	OpenAIRE aims to support the implementation of Open Access in Europe. OpenAIRE's three main objectives are to: • Build support structures for researchers in depositing FP7 research publications through the establishment of the European Helpdesk and the

	outreach to all European member states through the operation and collaboration of 27 National Open Access Liaison Offices; • Establish and operate an electronic infrastructure for handling peer-reviewed articles as well as other important forms of publications (pre-prints or conference publications). This is achieved through a portal that is the gateway to all user-level services offered by the e-Infrastructure established, including access (search and browse) to scientific publications and other value-added functionality (post authoring tools, monitoring tools through analysis of document and usage statistics); • Work with several subject communities to explore the requirements, practices, incentives, workflows, data models, and technologies to deposit, access, and otherwise manipulate research datasets of various forms in combination with research publications.
Interoperability objectives	• Create an international infrastructure for the storage and access of research output. Interoperability of services, standards and policies is an important component of this infrastructure.
Link	http://www.openaire.eu

Name	**Open Access Repository Junction (OA-RJ)**
Domain	Open Access Repositories
Timescale	2009-2011
Description	The project aims to assist open access deposit into, and interoperability between, existing repository services building and testing a deposit broker tool. Specific goal is to simplify the repository deposit workflow for multiple-authored journal articles.
Interoperability objectives	• To address interoperability problems currently faced by researchers who have written a multi-authored journal article from multiple institutions and grant-funding organizations. • To develop solutions to enable multiple deposit
Link	http://edina.ac.uk/projects/oa-rj/index.html

Name	**Open Archives Initiative (OAI)**
Domain	Open Archives
Timescale (start date)	2002
Description	The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication.
Interoperability objectives	• To promote interoperability standards to facilitate open access to a range of digital content • Interoperability framework for archives
Link	http://www.openarchives.org/OAI/OAI-organization.php

Name	Open Archives Initiative Object Reuse and Exchange (OAI-ORE)
Domain	Open Archives
Timescale (start date)	2006
Description	OAI-ORE is an initiative that aims to solve the problem of indentifying an aggregation of resources on the web. An aggregation is a group of related content, such as for example different formats of the same document, or distributed resources with multimedia types. OAI-ORE uses Resource Maps (in Atom, RDF/XML and RDFa) to describe the constituents or boundary of aggregations. This is done by providing aggregations with individual URIs and machine-readable data describing it, including information about the aggregation itself (like who created it) as well as information about the relationships between the resources in the aggregation. OAI-ORE can be expressed in several semantic formats such as RDFa, RDF/XML, turtle and n3. A powerful tool for digital libraries, OAI-ORE is being used by Chronicling America to provide linked data about aggregations of items such as the individual pages in an issue of a newspaper.
Interoperability objectives	• Providing a standard way to identify aggregations and expose the rich content in this aggregations to applications that support authoring, deposit, exchange, visualization, reuse and preservation.
Link	http://www.openlibrary.org/

2.4.5 Persistent Identifiers

As widely investigated in the WP22 of the APARSEN project[15], interoperability between persistent identifiers (PIDs) is one of the key challenge for guaranteeing persistent discoverability, accessibility and reuse of digital resources (Bouquet, Bazzanella, Riestra, & Dow, 2011). Recently, more awareness has been raised about this topic as shown by several initiatives focused on persistent identifiers interoperability[16] also with special attention to digital preservation issues.

In this section, we describe the past and current initiatives on this topic.

Name	PersID
Domain	Persistent Identifiers for digital publications and other resources
Timescale	2009-2011
Description	The project aimed to build a persistent identifier infrastructure for digital publications and electronic resources. The idea behind the PersID initiative is to provide an independent, flexible and trustworthy system of identifying resources

[15] See the D22.1, titled Persistent Identifiers Interoperability Framework, at http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework

[16] For example, the Seminar on Persistent Object Identifiers, the Hague, 14 and 15- 06-2011; "Links that lasts", Cambridge, 19-07-2012; "International Workshop on Interoperability of Persistent Identifiers Systems", Florence 13-12-2012.

	and making reliable links to them through implementation of an international standard system, the National Bibliography Number (NBN). The identifier system proposed by the PersID initiative is URN.
Interoperability objectives	• to implement a global resolver infrastructure, which is be able to resolve requests to local URNs but also to other Persistent Identifiers. • providing persistent identifiers as well as a transparent policy and technical framework for using them in the Internet.
Link	http://www.persid.org/index.html

Name	**DIGOIDUNA**
Domain	Digital Object Identifiers and Author Identifiers
Timescale	2011-2012
Description	DIGOUDINA is a study on identifiers for digital objects and authors, conducted by the University of Trento on behalf of the European Commission. The objective of the study is to support policy makers at European, Member State and research institution levels in assessing impacts and understanding the opportunities and challenges connected to managing digital identifiers within the context of scientific data e-infrastructures (SDIs), providing instruments to support decision makers on solutions that will have a long-lasting influence on scientific research and on the long term access, preservation and integration of valuable data and knowledge assets held within the sector.
Interoperability objectives	• Analysing the fundamental role of identifiers as enablers of value in e-infrastructures and presenting forward looking scenarios as examples of the benefits of a systematic usage of identifiers for digital objects and authors to locate and integrate information from multiple sources; • Reporting the results of the analysis of the Strengths, Weaknesses, Opportunities and Threats (SWOT) associated with establishing in Europe an open, dynamic and sustainable governance of e-infrastructure using identifiers for digital objects and authors; • Presenting the main challenges and recommendations which European Commission and other relevant stakeholders should address to develop an open, interoperable and sustainable e-infrastructure for locators of digital objects and identifiers of authors supporting scientific information access, curation and preservation.
Link	http://digoiduna.wordpress.com/ http://www.digoiduna.eu/

Name	**RIDIR Project**
Domain	Persistent Identifiers for digital repositories
Timescale	08/03/2007 08/10/2008
Description	The RIDIR project (Resourcing Identifier Interoperability for Repositories) is a project funded under the auspices of the Joint Information Systems Committee 'Repositories and Preservation' Programme. The aim of the project was to investigate the requirements for, and benefits of, the clear use of persistent identifiers in order to facilitate interoperability between digital repositories of different types.

Interoperability objectives	• To engage with the identifier and repository communities to understand better their requirements and highlight the benefits of the clear use of persistent identifiers in order to facilitate interoperability where required. • To develop and build a fully working demonstrator to showcase the findings of this engagement and demonstrate potential means for addressing the issues raised.
Link	http://www2.hull.ac.uk/discover/ridir.aspx

Name	**PILIN (Persistent Identifier Linking Infrastructure)**
Domain	Persistent Identifiers
Timescale	2006-2008
Description	Part of the ARROW initiative in Australia, it is a project which aimed to strengthen Australia's ability to use global persistent identifier infrastructure, particularly in the repository domain. A major focus of the project was the definition of a policy framework for managing identifiers including guidelines, requirements, instances of policy documents, a glossary, an ontology and a Service Usage Model.
Interoperability objectives	• Develop an abstract model for identifiers and their management • Develop shared, standards-based, persistent identifier management infrastructure • Support adoption of persistent identifiers and services • Plan for sustainable shared identifier infrastructure
Link	http://www.arrow.edu.au/PILIN.php

Name	**Knowledge Exchange initiative on persistent identifiers**
Domain	Persistent Identifiers
Timescale (start date)	2005
Description	The Knowledge Exchange (a collaboration of ICT/library organisations from Denmark, Germany, the Netherlands and the UK) has embarked upon an initiative to bring together parties that agree to a common approach to establish an international infrastructure that uses URN:NBN to provide permanent, and preferably open, access and reference-opportunities to unambiguously identified publications and other content when browsing the internet.
Interoperability objectives	• Providing guidelines for an international harmonized persistent identifier framework that serves the long-term preservation needs of the research and cultural heritage communities • Prototyping a metaresolver for URNs
Link	http://www.arrow.edu.au/PILIN.php

Name	**EPIC: European Persistent Identifiers Consortium**
Domain	Persistent Identifiers

Timescale (Start date)	2009
Description	It is a consortium of several partners (GWDG, SARA, DKRZ, CSC, CLARIN, DARIAH-EU) to provide a PID service based on Handle, http://www.handle.net/, for the allocation and resolution of persistent identifiers. Since the beginning of 2009 the service has been run by GWDG on behalf of the Max Planck Society.
Interoperability objectives	• Representing relations between scientific data and other resources • Long-lasting accessibility to scientific data • To allow new methods to reference the primary and secondary scientific data in order to name these data in a unique and timeless way like the ISBN numbers for books, which are permanent and citeable references to the related books.
Link	http://www.pidconsortium.eu/

Name	**Data Citation Standards and Practices Task Group**
Domain	Data citation standards and good practices
Timescale (Start date)	2010
Description	The Task Group aims to examine a number of key issues related to data citation, help coordinate activities in this area, and promote common practices and standards in the scientific community.
Interoperability objectives	• To address interoperability of datasets formats • To address interoperability of citation formats • To address interoperability of PIDs • Implementation of data citation standards and good practices
Link	http://www.codata.org/taskgroups/Tgdatacitation/

Name	**Linked Content Coalition**
Domain	Online copyright
Timescale	2012
Description	It is a global industry project that aims to develop a detailed reference model for linking items of content and rights on digital network. The final goal is to “create the framework for a fully interoperable and fully connected standards-based communications infrastructure so that businesses and individuals can manage and communicate their rights more effectively online”. A major work stream of LCC is that of persistent identification of entities within the Rights Reference Model which include (but are not limited to) objects, authors and institutions. The model is a development of the indecs content model (http://en.wikipedia.org/wiki/Indecs_Content_Model).
Interoperability objectives	• Create the framework for a fully interoperable and fully connected standards-based communications infrastructure so that businesses and individuals can manage and communicate their rights more effectively online.
Link	http://www.linkedcontentcoalition.org

Name	CrossRef Joins STM-DataCite Statement
Domain	Linkability and Citability of Research Data
Timescale	June 2012
Description	In June 2012, DataCite and the International Association of STM Publishers (STM) issued a joint statement on the Linkability and Citability of Research Data (http://www.stm-assoc.org/2012_06_14_STM_DataCite_Joint_Statement.pdf). In August 2012, CrossRef joined and support this statement and the best practices for data it recommends. This is a further step toward the commitment to the interoperability of CrossRef and DataCite's services which are based on the Digital Object Identifier (DOI) System, recently approved as an ISO Standard (**ISO 26324:2012, *Information and documentation – Digital object identifier system***). Specifically, CrossRef encourages publishers to use DataCite DOIs to link to data sets referenced in the published literature, and encourages authors of research papers to use CrossRef DOIs to link from data deposited in DataCite repositories to the published articles that draw on that data. CrossRef and DataCite are also collaborating on joint services, such as DOI Content Negotiation (http://crosscite.org/cn/), to enable publishers and data repositories to automatically interlink their content.
Interoperability objectives	• Interoperability of CrossRef and DataCite's services which are based on the Digital Object Identifier (DOI) System.
Link	http://www.stm-assoc.org/2012_06_14_STM_DataCite_Joint_Statement.pdf

Name	ORCID
Domain	Open Research and Contributor ID
Timescale	June 2012
Description	ORCID aims to standardize research identification solving the name ambiguity problem in research and scholarly communications through the implementation of a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID and other current researcher ID schemes. These identifiers, and the relationships among them, can be linked to the researcher's output to enhance the scientific discovery process and to improve the efficiency of research funding and collaboration within the research community. An author can either create a new ORCID ID or import profile information from an existing profile system (e.g., Scholar Universe, Researcher ID, Scopus, REPEC). Once an author has an ORCID ID they can export this profile information from ORCID to relevant stakeholder systems.
Interoperability objectives	• Interoperability by standardization among several Authors Identification system
Link	www.orcid.org

Name	OKKAM
Domain	Unique global identifiers for digital and non-digital entities
Timescale	2008-2010
Description	The overall goal of the OKKAM project was enabling the Web of entities, a global information space for publishing and managing information about digital and non-digital entities, including digital objects, people, locations and many others. The project aimed at providing an infrastructure, named Entity Name System, for assigning and managing unique identifiers for entities of any kinds, and tools on top of it to foster the systematic re-use of unique entity identifiers for information integration and entity-centric search and retrieval.
Interoperability objectives	• The ENS can be seen as an infrastructure capable of providing data integration across multiple domains. For example combining Semantic Web technologies and unique persistent identification of resources it is possible to support entity level integration of information distributed across independent systems. • To facilitate entity identifiers reuse across platforms, applications and domains. • Providing a technological infrastructure to enable the development of an interoperability framework for PIDs (alternative IDs index, resolver index).
Link	www.okkam.org

2.4.6 Semantic Interoperability and Linked data

As we have described in Section 2.3, the Linked Data initiative aims at interlinking heterogeneous data and metadata in order to facilitate the discoverability of and access to Web resources in a unified way. It is not surprising that Linked Data appeared to be the perfect paradigm to approach the interoperability problems faced by libraries, such as data interoperability, unified data access and interconnecting data silos.

Libraries have been adopting the Linked Data approach from the very beginning of the initiative.

Some examples of the experience of libraries with Linked Data are:

- Linked Data at the California Digital Library which is using principles of Linked Data and the REST architecture[17] in the implementation of its digital repository micro-services known collectively as Merritt. Linked Data is used for integration/coordination of functionally distinct curation services.
- Linked Data and LIBRIS, the Swedish National Union Catalogue maintained by the National Library of Sweden (http://libris.kb.se/)
- Linked Data at the Library of Congress (see the BIBFRAME initiative below)

A list of Linked Data Datasets (mostly bibliographic) is reported in the Appendix I.

[17] REST, Representational State Transfer, is a style of software architecture for distributed systems.

In this section we will focus on some relevant initiatives, which have specifically adopted the Linked data framework to face problems of interoperability related to digital preservation issues in the library context.

Name	**Linked Data at the Library of Congress: The Bibliographic Framework Initiative** (BIBFRAME)
Domain	Library
Timescale	2011-2013
Description	A major focus of the project is to translate the MARC 21 format to a Linked Data (LD) model while retaining as much as possible the robust and beneficial aspects of the historical format.
Interoperability objectives	• A flexible mechanism to accommodate future cataloguing domains, and entirely new use scenarios and sources of information • Using the Web as an architectural model for expressing and connecting decentralized information • Social and technical adoption outside the Library community • Social and technical deployment within the Library community • Exploiting previous efforts in expressing bibliographic material as Linked Data • Application of machine technology for mechanical tasks • Exploiting previous efforts for 64itabilit bibliographic information in the library, publishing, archival and museum communities • Defining a common method of bibliographic information transfer
Link	http://www.loc.gov/marc/transition/ http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf (the document offers a high-level view of the BIBFRAME model)

Name	**W3C Library Linked Data Incubator Group**
Domain	Library
Timescale	2010-2011
Description	The mission of the group is “to help increase global interoperability of library data on the Web, by bringing together people involved in Semantic Web activities — focusing on Linked Data — in the library community and beyond, building on existing initiatives, and identifying collaboration tracks for the future.”
Interoperability objectives	• Leading a shared standardization effort within the library community around (Semantic) Web standards
Link	http://www.w3.org/2005/Incubator/lld/ http://www.w3.org/2005/Incubator/lld/wiki/UseCases (A use-case document that describes a number of real-world use cases, case studies, outreach and dissemination initiatives targeted to the library community)

Name	**The Open Library Project**
Domain	Library
Timescale (start date)	2006

Description	It is an online project intended to create "one web page for every book ever published". It is a project of the non-profit Internet Archive and has been founded in part by a grant from the California State Library and the Kahle/Austin Foundation. The output of the project is a large bibliographic database with metadata for books.
Interoperability objectives	• Creation of a quantity of linked open data for bibliographic items, from a variety of sources and usable in web applications. • The Linked Data technology can be used to easily reference specific manifestations in the OL data. This facilitates the access to different manifestations (i.e. editions) of the same work. • To provide a unique interface to access to a full text version available at the Internet Archive
Link	http://www.openlibrary.org/

Name	**LOCAH (Linked Open Copac and Archives Hub) Project**
Domain	Archives Hub
Timescale (start date)	2010-2011
Description	The LOCAH project aims to make records from the JISC funded Archives Hub service, and records from the JISC funded Copac service available as Linked Data. The aim is to provide persistent URIs for the key entities described in that data, dereferencing to documents describing those entities.
Interoperability objectives	• Aggregation and merging of Hub data by creating links between the Hub and other data sources like DBPedia, and the BBC, as well as links with OCLC for name authorities and with the Library of Congress for subject headings. • Allow free and flexible exploration of data enabling users to make new connections between entities (subjects, people, places, organizations…).
Link	http://archiveshub.ac.uk/locah/2010/07/23/locah-project-aims-objectives-and-final-outputs/

Name	**VIAF (Virtual International Authority File) as Linked Data**
Domain	Library
Timescale (start date)	2000
Description	The VIAF initiative, implemented and hosted by OCLC, is a joint project of several national libraries and transnational library agencies with the aim to linking widely-used authority files (i.e. 20 national-level name authority files) and making that information available on the Web. Now available as Linked Open Data (LOD), VIAF is leveraged by freebase.com and other agencies and services.
Interoperability objectives	• Providing a single authority service combining multiple library authority files
Link	http://viaf.org/

Name	**STITCH: Semantic Interoperability To access Cultural Heritage**
Domain	Cultural Heritage
Timescale	2006 (start date)
Description	The main goal of the project is to develop theory, methods and tool to promote semantic interoperability for metadata vocabularies through the building of semantic links between the vocabularies used for metadata. The project aims at investigating ways to represent metadata and vocabularies in RDF/OWL formats, which allow the building of resources linked at a semantic level. The starting point is the ontology mapping approaches of the Semantic Web. STICH is a CATCH project (See below)
Interoperability objectives	• through ontology mapping, creating the background knowledge required for accessing distributed repositories (both within organizations and on the Web) using different vocabularies.
Link	http://www.cs.vu.nl/STITCH/

Name	**CATCH: Continuous Access to Cultural Heritage**
Domain	Computer Science, Cultural Heritage and Humanities
Timescale	2005 onwards
Description	Since 2005 CATCH finances teams which focus on improving the cross-fertilisation between scientific research and cultural heritage. The teams consist of a PhD student, a post-doc and an IT programmer. In the light of transferability and interoperability, the research teams execute their research at the heritage institutions, according to the laboratorium extra muros formula. Currently CATCH is financing fourteen research projects conducted in twelve cultural heritage institutions. (For an overview of the CATCH projects, see: http://www.nwo.nl/nwohome.nsf/pages/NWOP_6CCC3L_Eng
Interoperability objectives	• All CATCH project use advanced computer science technology to improve the access and interoperability of cultural heritage collections. Humanities scholars form the most important user group of these technologies.
Link	http://www.nwo.nl/en/research-and-results/programmes/Continuous+Access+To+Cultural+Heritage+%28CATCH%29

Name	**CEDAR (Linked Open Census Data)**
Domain	Computational Humanities
Timescale (start date)	2011
Description	CEDAR takes Dutch census data as its starting point to build a semantic data-web of historical information. With such a web, it will be possible to answer questions such as: - What kind of patterns can be identified and interpreted as expressions of regional identity? - How can patterns of changes in skills and labour be related to technological progress and patterns of geographical migration? - How can changes of local and national policies in the structure of communities and individual lives be traced?

Interoperability objectives	The CEDAR project applies a specific web-based data-model – exploiting the Resource Description Framework (RDF) technology– to make census data inter-linkable with other hubs of historical socio-economic and demographic data and beyond. Pattern recognition appears on two levels: first to enable the integration of hitherto isolated datasets, and second to apply integrated querying and analysis across this new, enriched information space. Data analysis interfaces, visual inventories of historical data and reports on open-linked data strategies for digital collections will be some of the results of this project. The project will also produce generic methods and tools to weave historical and socio-economic datasets into an interlinked semantic data-web.
Link	http://cedar-project.nl/ (and http://ehumanities.nl/computational-humanities/)

Name	**LUCERO (Linking University Content for Education and Research Online)**
Domain	e-Science open data
Timescale	2011 (one year project)
Description	Project at the Open University (also funded by JISC) is investigating and prototyping the use of linked data technologies and approaches to extract, interlink and expose data available in various institutional repositories of the University and make it available openly for reuse. LUCERO is working on exposing a number of Open University data sets as Linked Data, including: Course Information Research Publications recorded in the Open University Research Online (ORO) repository People information (specifically OU staff) Podcasts Course material metadata – this is descriptive information from the Open University library catalogue covering books and other media (e.g. DVDs, CDs) but excluding online material A number of Open University research and teaching material resources.
Interoperability objectives	• To integrate linked data technology in a sustainable way to support the research and educational activities of a Further or Higher Education organisation
Link	http://lucero.open.ac.uk/

2.4.7 Semantic access to Earth sciences resources

This section falls within the activity of Task2530 of project's work package 25, as described within Section 1.2.

We now introduce the reader to ontology related activities in the domain of Earth sciences, as most recently investigated at the European Space Agency.

Name	**OTE: Ontology and Terminology for Earth Observation**
Domain	Earth Observation Ontology for Application Domain, controlled vocabulary, web services for ontology discovery
Timescale	2007 (start date)

Description	The ESA project Ontology and Terminology for Earth Observation (OTE) aimed to explore the possibility to spread the use of EO products by making it easier for final users to identify EO products relevant for their needs. For this purpose some application domains have been analysed and conceptualized, and EO products have been connected to the concepts that were identified. Mainly the "Marine and Coastal Environment" domain has been studied, but also some concepts from "Ice and Snow" and "Fire Risk" have been added. The resulting Knowledge Base has then been made available through web services and a web application.
Interoperability objectives	The objectives of the OTE project were: • to set up a prototype knowledge based system capable to record, in the form of mappings between ontologies, the connections between application specific terms from one test application domain (namely the "Marine and Coastal Environment") and EO specific terms used to describe EO resources (such as for instance mission, sensor, etc.); • to produce a demonstrative web application. This web application is capable of accessing the knowledge based system through web services and present the users with the proper selection of EO resources (and all related available information) with respect to the application terms/concepts the user is interested in. • to investigate about the possibility to host such a demonstrator into external web data repository
Link	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=OTE+Project

Name	**OTEG: Open Access Ontology for GMES Space Component**
Domain	Earth Observation Ontology for Application Domain, controlled vocabulary, web services for ontology discovery, GMES, GSCDA
Timescale	2008 (start date)
Description	Starting just after the prototype implementation of OTE, the OTEG project was intended to revise and expand the results of OTE. The purpose of OTEG was to implement, validate with the support of experts, and make openly available a full scale demonstrator. Furthermore OTEG added to the past achievements of the OTE project the implementations satisfying new requirements related to the initial GSCDA version. In 2008, a first set of services related to the GMES initiative became operational. At that time, ESA presented a series of new requirements concerning the development of the knowledge-based system to access EO data in an affordable way for potentially interested users who are not expert in the EO domain. The Global Monitoring for Environment and Security (GMES) is a European initiative for the implementation of information services dealing with environment and security. The initiative is led by the European Union (EU) in partnership with the European Space Agency (ESA), and aims at combining ground and space-based observations to develop an integrated environmental monitoring capability. GMES will therefore provide accurate, up-to-date and globally-available information on an operational basis to European, national, regional and local entities, enabling them to develop services and applications related to land, sea/ocean and atmospheric monitoring as well as to emergency response and security.

	The GSCDA (GMES Space Component Data Access) in particular comprises the semantic architecture intended to enable and ease the access and delivery of EO resources, and is part of the GMES Space Component (GSC), i. e. the space infrastructure that ESA is to develop and deliver within the GMES initiative.

Interoperability objectives	Within this frame, the main objectives for the OTEG project, building on top of the results achieved in the previous OTE project, were to develop:
	• a GSCDA Ontology extending the ones resulting from the OTE project, to support the GSCDA Application Domains and the GSCDA EO Domain (GSCDA EO resources);
	• an overall semantic architecture (the GSCDA Semantics), accessible via web application (and whose elements are downloadable via ftp), composed of:
	    o a Multi-domain Vocabulary,
	    o a Multi-domain Thesaurus,
	    o two Application Domain ontologies (and respective terminologies),
	    o and a Multi-domain Thesaurus augmented with a GSCDA Taxonomy (i.e. the mapping between application domains and EO resources);
	• User Client(s) for textual search and 3D graphical navigation of the above GSCDA Semantics

Link	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=OTEG+Project http://gmesdata.esa.int/OTE/navigateInfoDomain

Name	**SEPR: Semantic Rules for Automated ENVISAT Products Identification**
Domain	Earth Observation Ontology for Application Domain, controlled vocabulary, web services for ontology discovery, GMES, GSCDA, Semantic Reasoner
Timescale	2009 (start date)
Description	SEPR introduces some important novelties in the semantic architecture, in particular concerning the way mappings between application domains and EO resources are established. In addition, SEPR targets a full covering of the five core information domains considered in the GMES initiative.
Interoperability objectives	The main objectives of SEPR are: • to extend the application domains considered, so as to encompass all the five core GMES domains (namely “Marine environment”, “Land monitoring”, “Atmosphere monitoring”, “Emergency” and “Security”); • to improve the definition of mappings between application domains and EO resources by introducing an intermediary level of description of the EO resources that makes more apparent and explicit the reasons why a data product is relevant for a given web application term • To set up a dynamic link between application domain thesaurus and EO resource. Every application term point to a keyword which can be searched within an ISO compliant catalogue (e.g.: GeoNetwork)
Link	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=SEPR+Project

Name	**KLAUS: Knowledge-based EO Demonstrator for Land Use Management**
Domain	Processing component, distributed programming environment
Timescale	2009 (start date)
Description	The KEO Demonstrator with Models for Land Use Management (KLAUS) is an improvement and extension of the prototype Knowledge-based Earth Observation (KEO) project, which is an ESA project for a distributed component-based programming and processing system for EO data. The KLAUS project aims to define and evaluate some operational scenarios which exploit Image Information Mining and other processing services provided by the KEO system, powered by semantic annotation (S. D'Elia, 2004 and S. D'Elia et al., 2008).
Interoperability objectives	One of the objectives of KLAUS is to update and possibly modify the ontology search originally available in KAOS (the client application to KEO). The requirement concerning the Semantic architecture in KLAUS is an ontology system analogous (and very similar) to the one developed for OTEG and SEPR projects, accessible through an API (web service or Java library) and allowing for the following operations: • Retrieval of the Application Terms relevant to the terms entered as (free text) search keys, by looking into the Multi-Domain Vocabulary (this last has to be precisely an updated version of the Multi-Domain Vocabulary of OTEG / SEPR); • Tagging of new Processing Components by attaching to them some Application Term as labels, taken from the list of existing Application Terms recorded in the Multi-Domain Thesaurus; • Retrieval of all the Processing Components tagged with a given Application Term; • Listing of all the Application Terms in the Multi-Domain Thesaurus. The requirements have been satisfied by producing a couple of web services, named KEO Terminology Service and the KEO Ontology Service, by means of which the KEO system can offer semantic search capabilities: • Processing Components tagging with Application Terms. • Identification and ranking of relevant Processing Components by searching for a list of user terms. • Selection of a processing within KEO that is one of the Processing Components resulting from the search.
Link	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=KLAUS+Project

Name	**RARE: Rapid Response Support Server**
Domain	ISO, OGC, HMA, GMES, standardisation, catalogue, ontology, semantic reasoned,
Timescale	2010 (start date)
Description	The RARE project is a further evolution on same line of work of OTEG and SEPR. RARE introduces significant novelties including: • Web-based user interface driven by application terms (domain specific, i.e. land, ocean and atmosphere monitoring, emergency response and security)

	users are familiar with;
	• Application domains covering GMES thematic domains;
	  ○ technique able to resolves:
	  ○ terms searched by the users to application terms;
	  ○ location names to geographical coordinates (using gazetteers);
	  ○ application terms, geographical coordinates and time constraints to categories of EO products;
	  ○ EO product categories to actual services.
	• Access to any type of geo-localizable resources: pictures, maps, features, 71itability, processors;
	• Interfaces with existing catalogues and services using standardized protocols (OGC, ISO), as described within the Heterogeneous Mission Accessibility Cookbook (See reference section);
	• Interfaces with a client application to visualize the data.

Interoperability objectives	The objective of RARE is to build a distributed software system accessible through a Web-based user interface that allows searching for EO-related resources such as satellite images, maps and geo-localized features (e.g. 71itability and points of interests) using application domain terms, that is, using the terminology any potential user is accustomed with. RARE will provide a centralized service that interfaces with a number of on-line resources that were not integrated in the past. These resources include: • a terminology service supporting the navigation facility between related application terms; • a query analyser that augments the knowledge with unforeseen concepts and relations extracted from the Internet; • gazetteers used to resolve place names; • various reasoners mapping the application terms selected by the users to product categories; • a centralized catalogue service that collects the properties of on-line resources registered in any remote catalogue compliant to ISO/OGC standards (e.g.: CSW ISO AP, CSW ebRIM EOP/Sensor-ML/CIM, OpenSearch, (EO-)WCS, (EO-)WMS, WFS, etc…) RARE will hide as much as possible the technical information related to the resources themselves: the system determines, based on elaborate mapping rules defined with the help of domain experts, which categories of resources are valuable in regards to the application terms entered by the users. This mechanism greatly reduces the time and the knowledge needed to search for, and obtain, the right resource for the right problem. It brings the advantages of EO-derived resources into the users domain of interest with minimal effort.
Link	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=RARE+Project http://wiki.services.eoportal.org/tiki-index.php?page=RARE%20Project

Name	**SMAAD: Semantic Web Mediated Access Across Domains**
Domain	ISO, HMA, OGC, standardisation, catalogue, ontology, semantic metadata annotation

Timescale	2010 (start date)
Description	The main objective for ESA's SMAAD project is to support the semantic discovery of EO resources. It includes the easy linking of metadata keywords from different domain ontologies and thesaurus. It also enables the services used in ESA's HMA project to support these annotations. In case the standards are lacking some functionalities to reach this goal, they will be amended and promoted to the standardisation organisation. The project is currently running: a prototype will demonstrate the workflow from the semantic annotation of metadata to the discovery of the resources using ontologies of other domains.
Interoperability objectives	The most relevant interoperability objectives of the Semantic-web Mediated Access Across Domains (SMAAD) project are as follows: • INSPIRE/GMES Discovery Service Interoperability, permitting INSPIRE and GMES community users to discovery resources collected both on INSPIRE and GMES catalogues. This scenario relates to the use of the INSPIRE Conformance Class of the CIM EP for CSW defined in [OGC 08-197r1] to allow for the translation of catalogue requests and responses between INSPIRE and EO preferred catalogue protocols. Both the ISO AP [OGC 07-045] catalogues used by the INSPIRE Community and the CIM EP [OGC 07-038r3] catalogues used by the EO Community contain the same ISO Service and Collection metadata and could, therefore, be of interest to both communities for extended cross-community Resource discovery. This becomes possible if appropriate ISO AP to CIM EP and CIM EP to ISO AP transformation modules/services are used. • The client will permit the user to discovery resources registered in remote catalogues catalogue compliant to ISO/OGC standards (e.g.: CSW ISO AP, CSW ebRIM EOP/Sensor-ML/CIM, OpenSearch, (EO-)WCS, (EO-)WMS, WFS, etc…) • Ontology mediation: the discovery will be supported by GMES (RARE outcome) and GEMET ontology, addressing land monitoring domain. The system will provide the user a mediator module, permitting to get access to GEMET resources by using GMES terminology and vice versa.
Link	http://wiki.services.eoportal.org/tiki-index.php?page=SMAAD

Name	**GEOSS: Global Earth Observation System of Systems**
Domain	Global Earth Observation System of Systems, GEOPortal, ontology, interoperability, open access
Timescale	2005 (start date)
Description	The GEOSS 10-year implementation plan specifies that one of the functional components of GEOSS is to “exchange, disseminate, and archive shared data, metadata, and products”. In order to implement a functional component for data exchange the implementation plan specifies the following data sharing principles for GEOSS: • “There will be full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation. • All shared data, metadata, and products will be made available with

	minimum time delay and at minimum cost. • All shared data, metadata, and products free of charge or no more than cost of reproduction will be encouraged for research and education."
Interoperability objectives	The challenge for GEOSS is in facilitating the lossless exchange of data between heterogeneous systems whilst ensuring that an integrated system is able to fulfil the needs of Communities of Practice. This brings to the fore issues of both syntactic and semantic interoperability. Syntactic variability is inherited from the reuse of contributed components. Whereas, semantic variability comes from the multidisciplinary nature of the stakeholders and beneficiaries of GEOSS. There are several kinds of ontology, taxonomies, thesauri, and gazetteers used within the GEO Communities of Practice and standards adopted by GEOSS.A subtask has been established within GEOSS to review semantic interoperability between earth observation data in GEOSS. So far the subtask has produced a prototype ontology registry based on Semantic MediaWiki. The ontology registry will become a component of GEOSS Infrastructure. The subtask has also produced a prototype digital gazetteer which will support geocoding services within GEOSS. The registry will also include data models from registered international standards. So far the ontologies and taxonomies registered in the GEOSS Standards Registry include the: WMO Manual on Codes contains WMO standards for the representation of information, relating to weather, climate and water. TDWG LSID Vocabularies containing concepts used in the Life Sciences and biodiversity research.
Link	http://www.earthobservations.org/geoss.shtml

Name	**Use of semantics by European Environment Agency**
Domain	Shared environment, Environment, semantic
Timescale	2010 (start date)
Description	The European Environment Agency (EEA) implemented the principles of SEIS (Shared Environmental Information System) in its Report-net infrastructure. EEA designed an application which is called "Content Registry" (CR). It works as a search engine and web crawler and respects the principle that the (structured) data is stored where it is produced. It also stores data centrally and re-harvests the data from the place they are produced frequently. The data format adopted is RDF. Also XML data is harvested and converted to RDF using XSLT style sheets. EEA has also published a Web site called RDFdata to store the reference data in RDF format with the intention of upgrading the EEA data service with the same functionality in the future. In the SOER (European Environment State and Outlook Report 2010) 2010 project, EEA is also planning to use the above principles to publish data using LinkedData principles with RDF. The feed format is defined in http://svn.eionet.europa.eu/projects/Zope/wiki/SOERFeedSpec.
Interoperability objectives	Federated catalogues Harvesting of metadata from remote catalogues Standardised languages and interfaces
Link	http://ec.europa.eu/environment/seis/

Name	FP6 InterRisk
Domain	ISO, Risk Management, ontology
Timescale	2005 (start date)
Description	InterRisk addresses the need for for better access to information for risk management in Europe, both in cases of natural hazards and industrial accidents. The overall objective is to develop a pilot system for interoperable GMES monitoring and forecasting services for environmental management in marine and coastal areas. The InterRisk pilot will consist of an open system architecture based on established GIS and web services protocols, and the InterRisk services to be implemented for several European regional seas. The InterRisk pilot system and services will be validated by users responsible for crisis management in case of oil spills, harmful algal blooms and other marine pollution events, in Norwegian, UK/Irish, French, German, Polish and Italian coastal waters. Sample products from the different services are shown under Services. Direct links to these services will be included, once they have been put in operation.
Link	interrisk.nersc.no

2.4.8 Other domain-specific interoperability initiatives

Name	EpSOS
Domain	eHealth
Timescale	2008-2013
Description	epSOS is the main eHealth interoperability project co-funded by the EC. The project aims to design, build and evaluate a service infrastructure to enable cross-border interoperability between Electronic Health Record systems and e-prescription services in Europe. For the first time, mobile EU citizens will have the opportunity to use cross-border eHealth services when seeking healthcare abroad in one of the participating 23 epSOS pilot countries.
Interoperability objectives	• to develop feasible cross-border eHealth services • developing a practical eHealth framework and ICT infrastructure that enables secure access to patient health information among different European healthcare systems
Link	http://www.epsos.eu/home.html

Name	ISA
Domain	eGovernment
Timescale	2010-2015
Description	ISA is an European Commission programme to foster interoperability, sharing and re-use between European public administrations.
Interoperability objectives	• Developing interoperability solutions for facilitating efficient and effective cross-border electronic collaboration between European public administrations. One of these services has been recently launched to make

	easier to find and re-use specifications that have been used to develop information systems for governments (http://joinup.ec.europa.eu/catalogue/all). • Developing an Interoperability Architecture cluster aiming to further align cross-border and cross-sector IT infrastructures that are already available to ensure that the governmental information is exchanged in a secure and trustworthy manner.
Link	http://ec.europa.eu/isa/

3 INTEROPERABILITY SCENARIOS AND CHALLENGES

To frame the discussion around interoperability and start to identify interoperability objectives, recommendations and services, we collected a set of interoperability scenarios and challenges. Some scenarios have been directly extracted from other deliverables of the APARSEN project (and we will refer to them for more details) and other sources. The scenarios are organized into different clusters pertaining different areas of the digital preservation landscape or specific domains (e.g. Earth Science). Since it is out of the purpose of this document to present an exhaustive list of interoperability scenarios for each area, we present here some relevant examples for each area and provide an overview of the key general challenges, which summarize and capture the main interoperability issues encountered in the analysis of each domain.

Since we aimed to use scenarios as operational tools for the interoperability assessment of the digital preservation landscape and to identify and prioritize the gaps described in Section 5.1, we asked partners who provided the scenarios to evaluate three dimensions, i.e. 1) the current situation about the raised issue, 2) the importance/impact of the issue, 3) the level of difficulty to address the problem. Optionally, they were asked to give a short justification of their evaluation quantified according to the following scales.

1. **Current situation**:
    a. Very Bad
    b. Bad
    c. Fair
    d. Good
    e. Very Good

Short justification of the answer (optional)

2. **Importance** (i.e. impact):
    a. Low
    b. Medium
    c. High

Short justification of the answer (optional)

3. **Level of difficulty** (to achieve):
   a. Easy
   b. Fair
   c. High

Short justification of the answer (optional)

For those scenarios that have been extracted from other sources, the evaluation has been partially provided (in the sense that only part of the scenarios have been evaluated or have been evaluated on a subset of dimensions) based on the experience and knowledge of the WP partners.

3.1 PERSISTENT IDENTIFIERS (PIDs) INTEROPERABILITY, KNOWLEDGE DISCOVERY AND CITABILITY

A number of scenarios and uses cases have been reported in the D22.1 of the project and we refer to this document for a comprehensive description of the interoperability issues regarding PIDs. However, since the connection with the WP22 is particularly relevant for the present work due to the pervasive impact of using PIDs in many domains of digital preservation, we report here some scenarios that have important implications for designing digital preservation actions and solutions.

Knowledge discovery and data integration through PIDs

Scenario 1:
Ellen is a researcher at the Childhood Bilingualism Research Centre of Hong Kong and she is conducting a psycholinguistic study on how children process words and sentences in Cantonese, English and Mandarin. During the review of the literature regarding this topic on a psychology database, she finds a paper published by a researcher of the University of Ottawa about how bilingual babies learn new words differently than monolingual babies. She finds the results reported in the paper very interesting for her study and decides to discover other works by the same author or other authors with similar research interests. She wants also to collect more information about the research activities and projects in which the author is (or has been) involved, the network of his collaborators and the institutions where he had previously worked in order to explore possibilities for research collaborations in future. A keyword-based research on a Web search engine shows that there are several authors and people with the same or similar name to that of the searched author and his name is recorded in several different ways. The author CV available on the Ottawa University Web site lists 3 main institutions where the author worked in the past and the list of publications seem incomplete and out of date. At a first glance, it seems to Ellen that finding exhaustive and integrated information about the author is not a trivial task. Fortunately, she remembers of a new knowledge discovery system for e-Science, called XYZ, built on an interoperability infrastructure for PIDs for authors, digital objects, organizations and other kinds of entities. Looking at the paper she is interested in, she notices that the paper reports such an identifier for the author and another identifier to identify the paper itself. Using the resolution system offered by XYZ and the author ID, Ellen is able to disambiguate the name of the author and find all the relationships between the author, his publications and scientific contributions. In particular, she finds the complete lists of the papers published by the author and his co-authors, (including open access papers), the links to papers on the same research topic and links to the datasets that are made available by the research project funding the study and an integrated profile of the research activities, collaborations and citation metrics about the author. She also finds many

references to track online contributions about the research topic, such as scientific blogging and community efforts.

Scenario 2:
Nachiket is an Indian student and he is looking for a PhD position in molecular biology in Europe. On the European Commission Web site, Nachiket finds the link to a new entity-based system that provides information to people seeking to advance their education and careers by moving to other countries. Exploiting the potentialities of unique identifiers, the system is the entry point to a wealth of integrated information from many different sources. Starting from a job vacancy announcement, the system allows obtaining information about all the mentioned entities in it such as people, locations, institutions, groups, programs and many others entities in some way related to these entities. George finds an announcement for a PhD position at the University of Warsaw and he starts to navigate the system to retrieve all the information he is interested in. In addition to what it is reported in the announcement, he wants to know more information about the project and the people involved in it, but also more practical information on living and relaxing in Warsaw. Thanks to unique identifiers for people, institutions, groups and projects, the system integrates all the information available on the Web referring to these entities and returns results tailored to the specific requests of the user. For example, George is able to access to the CV and the complete list of publications of the people working in the same project, he can know all the funding details of the project and have more information about the other institutions involved. The digital identifier of the university allows George to find all the information to apply for university housing, legal information about the job position in the foreign country, procedure for VISA application and so on. Through the identifier of Warsaw, the system returns a map of the city that can be searched using many different kinds of categories, such as best-rated restaurants, dog-friendly or free-admission amenities. An interface to explore the public transportation system of the city, a booking service to incoming events and movie tickets are also provided with a simple click. What a surprise for George to discover that all the questions that are coming in his mind find an answer without the need to recur to external Web sites!

Challenge (scenario 1 and 2): given a persistent identifier, a user wants to retrieve information about the identified entity (digital object, author, organization and so on) and other related entities. Since this information is usually distributed across many different systems, which usually adopt different identification schemes to identify the same entity, the interoperability challenge is to enable the access to this distributed information through a single entry point which make different PI solutions interoperable.

1. **Current situation**:
    a. Very Bad
    b. Bad*
    c. Fair
    d. Good
    e. Very Good

Short justification: The current situation is bad, but there are some growing initiatives and systems (e.g. ORCID for authors, OKKAM ENS for any kind of entity), which have started to address the PIDs interoperability challenge. However, these solutions are not widely adopted and this represents a big obstacle for their bootstrapping and their global cross-boundary effectiveness.

2. **Importance** (i.e. impact):
    a. Low
    b. Medium
    c. High*

Short justification: The relevance of the challenge has become very high in the last years with the growing of digital material about the same entities distributed across heterogeneous sources and available through the Web (for example as linked data). The distributed and heterogeneous nature of the information to be accessed poses new challenges for digital preservation approaches.

3. **Level of difficulty** (to achieve):
    a. Easy
    b. Fair*
    c. High

Short justification: a technical solution to the challenge is not very difficult to implement, but it is worth to say that the success of the solution is far to be a mere technical problem, concerning, economical, political and social dimensions. The very difficult task here is to build a community, which agrees on policies, responsibilities, funding models etc. beyond the technical solution.

Author Identifiers Interoperability (from D22.1)

Scenario 3
The premise here is that authors change organisation affiliation over time or may have multiple organisations (funding body, university, project) at one time and may therefore be assigned several disconnected author persistent identifiers.

a) An Archive is ingesting a data collection from a given author and adds that author to their local database with the persistent author ID provided. The Author ID is used to identify other related publications from the author for presentation to end-users. The Archive enters the Author ID into the Interoperability System and a number of possible matches are returned which might represent the same author. The Archive is able to contact the author directly during ingest negotiations and confirms which of the other Author ID's represent the same person. In this way the Archive can 'link' the IDs and the Archive's presentation of related publications is greatly enriched.

b) Ideally a mechanism exists for the Archive to submit the linked identifiers back to the interoperability system and the system can respond to future author queries with increased confidence. The Organisation ID of the Archive could be presented alongside the associated author IDs on subsequent searches to clarify the provenance of the correlation between the IDs.

Challenge: association of multiple persistent identifiers with a single person to enhance information integration from different sources and provenance assessment.

1. **Current situation**:
    a. Very Bad *
    b. Bad
    c. Fair

d. Good
e. Very Good

Short justification: as clearly shown by the survey on Persistent Identifiers systems conducted in the WP22 of the APARSEN project[18], there is a different level of maturity between the more advanced systems for digital objects and the gradually emerging solutions for authors. More than 50% of the participants reported that they don't adopt persistent identifiers for authors and contributors. This is partly explained by the relatively recent start of important initiatives (e.g. ORCID, ISNI, AuthorClaim), but also by the fact that many concepts and solutions devised for managing persistent identifiers for digital objects cannot be directly applied to authors. Authors are persons, and as such many legal and social constraints apply on what can be used (for what, under what conditions) in profiling authors and combining data about them. On the other hand, the description of an author is by definition dynamic, namely changes through time; this collides with a basic assumption of PIDs for digital objects, namely that the object identified through a persistent ID should not change (this is indeed one of the fundamental principles of ensuring integrity for a publication or a dataset).

2. **Importance** (i.e. impact):
   a. Low
   b. Medium
   c. High*

Short justification: Interoperability between author persistent identifiers may have a beneficial impact for many different stakeholder communities and for many digital preservation tasks. In Table 3 we report a list of expected benefits by stakeholder type.

Type of Stakeholder	Benefits
Individual Authors / Researchers	• to assemble their career profiles and lists of publications (even if the author has worked for several different institutions) • to claim their authorship • to find information (e.g. find all the papers of the same author or research group) • to find potential collaborators • to monitor their performance (citation metrics) or that of a competing research group; • to identify reviewers for an article • to get credit for their scholarly activities
Institutions: Universities, Research Institutes, Data centres	• to evaluate the scholarly activities through academic metrics • to assess candidates for promotion and tenure • to track the achievements of their researchers • to improve open access services to their repositories (e.g. to find all scientific authors of documents at any repository)

[18] Se D22.1 available at http://aparsen.digitalpreservation.eu/pub/Main/ApanWp22/APARSEN-REP-D22_1-01-1_8.pdf (p. 78)

Publishers	• to simplify the publishing workflow (e.g. submission, reviews…) • to integrate the data it into their commercial services
Libraries and Digital Libraries	• to match and link authority records • to improve knowledge discovery • to link published items to related datasets held by other institutions.
Funding Agencies	• to simplify the project submission workflow • to track the progress status of the research they funded • to make impact assessment of funded projects much faster and effective
Service and content providers	• to implement advanced knowledge discovery services, cross-referencing services, citation and statistics services.
Author rights organizations	• to protect intellectual property rights • to detect plagiarism
Governmental Institutions	• to produce indicators of scientific and technological production • to guide the implementation of policies to support research activities

Table 3: Expected benefits of interoperability between PIDs for authors by stakeholder type

3. **Level of difficulty** (to achieve):
   a. Easy
   b. Fair
   c. High*

Short justification: Author identification is a very complex problem, which involves many different stakeholders, which sometimes have different views and attitudes about the issues that need to be addresses. Therefore, devising a solution to the problem of interoperability for author identifiers is far from being a merely technical issue involving decisions about privacy, business models, openness, trust and many other issues. Raising awareness about the importance of using PIDs for authors is one of the key challenge for the success of the author identification initiatives.

Impact and Quality Assessment (from the DIGOIDUNA study)

Scenario 4

Stefano is a Project Officer who works for the European Commission. He is filling the Impact Assessment Form, which is part of the final evaluation report of a project he has been assigned. The project adopted the new EC policy on persistent identifiers, which means that any relevant entity connected to the project has a unique identifier, which is compliant with the pan European e-Infrastructure for research data on climate change. This means that the project itself has an identifier, together with every contributor (scientists, PhD students, developers, experts,...), every involved organization (universities, research institutes, companies, public bodies, ...). Any relevant entity is also geo-referenced. Stefano logs into the Project Impact Assessment portal and uses the project ID to collect all he needs to know to fill the form: the complete list of papers published or submitted for publication by the project team, a rich collection of metadata about them (type of publication, degree of inter-organizational collaboration, impact factors of journals, statistics on external and self-citations, and the like), a list of patents filed in the course of the project. He can browse the graph of EC funded projects and explore the history of past collaborations. Stefano can retrieve (and visualize on a map) other data, which prove the impact of the project in terms of jobs created within the participating partners, number and location of start-up companies and compute an overall impact indicator.

The same happens for every EU funded projects, which in the mid terms allows policy makers to make much better informed decisions on the assignment of funds and provide a much more complete account to member states governments and European citizens.

Challenge: to capture much richer relationships between ongoing and past projects and related entities (authors, institutions, collaborators…) to assess the quality and the impact of them across time.

1. **Current situation**:
    a. Very Bad
    b. Bad*
    c. Fair
    d. Good
    e. Very Good

Short justification: the lack of use of Persistent Identifier systems for authors and the fragmentation of the landscape of PIDs systems for digital objects make the impact and quality assessment issue very challenging. The situation is complicated by the fact many PIDs are used within territory boundaries. The usage of NBN is an example of PIDs that operate in practise within (part of) the European territory. The interoperability between PIDs for authors, digital objects and other kinds of entities would provide a route for reconciliation of identifiers for contents originating across territories. Currently, however, the benefits of managing high quality persistent identifiers at global and web-scale are not capable of being fully realised due to the fragmentation of the ecosystem.

2. **Importance** (i.e. impact):
    a. Low
    b. Medium
    c. High*

Short justification: Research production plays a crucial role in our economy, which in turn means that management and monitoring of funding and research information – connectivity of authoritative identifiers between the parties involved in provision and consumption of funding – is of crucial benefit over the research activity lifecycle and may have an important impact on the global economy.

3. **Level of difficulty** (to achieve):
    a. Easy
    b. Fair
    c. High*

Short justification: the fragmentation within the PIDs ecosystem as well as the need of identification of different kinds of entities (digital objects, authors, contributions, institutions) in order to address the impact and quality assessment challenge, make the integration task very difficult to be addressed.

Citability of Scientific Datasets

Scenario 5

A research group at the Beijing Genomics Institute (BGI) has recently worked in collaboration with the University Medical Center Hamburg-Eppendorf researchers to understand a new virulent strain of E. Coli causing an outbreak of food poisoning that has killed 18 people In Europe. Using the genomic technology, scientists have determined the infectious strain and revealed the mechanisms of infection, facilitating the development of measures to control the spread of this epidemic. To maximise the utility of the findings to the research community and help scientists around Europe to find a treatment for the infection, the Chinese research group decides to release the genomic data into the public domain.

Some months later, a researcher (or a reviewer) would like to examine the data used by Beijing group.

To this purpose she needs to access a description of the data that was used, the research protocol adopted, contributors, date and other information. Moreover she needs to re-use the data to compare the results from different experimental manipulations. The researcher wishes also to find if other data of the same type and from the same contributor are available, if articles co-authored by the same contributor have been published and other articles reporting results on the same dataset can be accessed.

Challenge: there are a number of interoperability challenges with this scenario.

- To identify, describe and access data that is referenced in scientific publications
- To link data to contributor
- To identify persons across boundaries: a person may have different roles as an author, a researcher, a data contributor. The identification need to be reused or linked to other systems such as those used by journals, data repositories, Linked Data
- To identify entities across systems which can use different vocabularies to classify them.

1. **Current situation**:
    a. Very Bad
    b. Bad
    c. Fair*
    d. Good
    e. Very Good

Short justification: growing initiatives in data citation have been started to address the challenges of making research data visible and accessible for the long time. DataCite is one of these initiatives which 1) support researchers to identify and cite research datasets, 2) support data centres by providing persistent identifiers for datasets, workflows and standards for data publication, 3) support journal publishers by enabling research articles to be linked to the underlying data. However, it is worth to say that the current situation is still characterized by the fact that the vast majority of research data are never shared and remain on the personal computer of the researcher or at best on the server of the research institution.

2. **Importance** (i.e. impact):
    a. Low
    b. Medium
    c. High*

Short justification: In recent years we have assisted to an exponential growth of research data. Most of the stakeholders within the research and scholar ecosystem, like researchers, publishers, funding agencies, libraries, agree about the benefits of having research data available, findable and reusable. Among the benefits we can include the following:

- To improve the availability, localization and reusability of research data.
- To link dataset to related publications.
- To make datasets legitimate and citable contributions.
- To make results of research verifiable and repurposed for future study.

3. **Level of difficulty** (to achieve):
    a. Easy
    b. Fair*
    c. High

Short justification: even though some initiatives have started to address the challenge some issue should be considered carefully such as, the level of granularity (at which level of granularity should data be made citable?), versioning (different identifiers for different updates of the dataset because it would harm the integrity of publications if the data they cited changed over time), access to descriptive metadata about the dataset (human or machine readable).

3.1.1 Generalized challenges about PIDs

1. **To provide a global resolution mechanism, which ensures that given an identifier of any kind the correspondent resource can be persistently retrieved and accessed. If the resource is not available any more, a matching resource if available (also from a different provider) should be linked.**
2. **To provide a unique interface to find integrated information across different systems about an identified entity (e.g. a paper) and related entities (related publications, authors, datasets…).**
3. **To create a collection from resources, that belong together (e.g. enhanced publications).**
4. **To associate multiple identifiers with the same entity (e.g. author) to enable the long term access to the entity or a description of it.**
5. **To locate all versions of a resource.**
6. **To find information about a resource's authenticity and availability.**
7. **To integrate metadata referring to the same resource from multiple sources.**
8. **To make citation and their relationships more explicit so that data can be accessed more easily, supporting re-use and verification and strengthen the link between the contributor and data.**
9. **To define a standard to uniquely identify datasets and manage them as separately citable items.**

3.2 METADATA INTEROPERABILITY AND LIFECYCLE MANAGEMENT

Vocabulary alignment

Libraries and other cultural heritage institutions adopt a variety of vocabularies and metadata element sets to describe items in their collections. Different vocabularies can be used within the same institution or in separate institutions to describe different collections (e.g. the main collections, special collections, multimedia collections) and there are usually no explicit links across vocabularies that indicate that two concepts or metadata fields have a similar meaning. This reduces the semantic interoperability between terms and concepts and raises some crucial problems:

- Resource discovery over different collections: Library communities develop discovery services for consortia with a geographical, subject, sector (public, academic), and/or domain (libraries, archives, museums) focus, either based on distributed searching or metadata aggregation. Trans-national consortia (e.g., Europeana) add a further layer of complexity of by including vocabularies in multiple languages.
- Metadata publishing, exchange and managing: the proliferation of local vocabularies reduces the efficiency of managing and publishing metadata and increases the need of vocabulary alignment and mapping.

Scenario 6[19]

Bob wants to search for books about a certain topic. Bob uses the catalogue system of his local library. Bob thus formulates an appropriate query using the category system and the thesaurus of his local library (e.g. his local library uses the category system of the German National Library and the German Schlagwortnormdatei).

Unfortunately, his local library has no books covering the searched topic. Bob thus also wants to search in the catalogues of further libraries which are participating in the interlending system. But some of the other libraries use another category system and thesaurus (e.g. the Dewey Decimal Classification and/or the Library of Congress Subject Headings). Thus, his original query has to be transformed in order to search in the catalogues of all libraries participating in the interlending system.

Challenge: Automatic mapping between different categorization systems and thesauri to implement search services for searching for resources that match a given query, over various catalogue systems.

Creating semantic links between archival collections and other (Web) sources[20]

Scenario 7

A researcher, Ella, is interested in people who have created, or who are significantly referred to in archives about the Second World War, but has been finding it difficult to bring her research materials together, as the archives of relevance are so widely distributed. She is particularly interested in ‘Sir

[19] http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Browsing_And_Searching_In_Repositories_With_Different_Thesauri

[20] http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_LOCAH

Winston Churchill' and wants to find archives created by him, or archives that have significant references to him. When Ella browses for 'Winston Churchill' on the Archives Hub, she discovers a fair scattering of Churchills as name entries – all apparently the same man, but it is difficult for her to get an overall sense of where these archives are held across the UK. Ella thinks it would be great if she could quickly see where all the materials on or by Churchill are held, to help her plan her research trips. Ella also thinks it would be great if the descriptions could include images, and other information about Churchill that she knows is available on Wikipedia.

Challenge: to make semantic connections between archival collections and other sources referring to the same entities (people, organisations, places and subjects).

Integrated metadata search interfaces across several providers for accessing digitized objects (from Europeana)

Scenario 8

Daniel is a graduate student in Computer Science at the University of Trento. He is looking for resources having to do with "Linked Data". He is interested not only in academic resources but also in events, blogs, online courses call for job positions and so on. He would also like to see videos (e.g. lectures or talk), or listen podcast on the topic. To find this information Daniel would have to consult a variety of data sources, each with a different system and interface, for each type of resource needed. Some resources found on the Web are not accessible, others are not relevant, and it is difficult to judge if the sources providing the contents are authoritative. What could help Daniel? Bob, Daniel's friend, suggests him an internet portal that acts as an authoritative interface to different types of content from different types of heritage institutions. Through the portal Daniel can search contextual information – or metadata – about digital objects related to the Linked Data topic. Once he finds what he is looking for, he can access to the full content of the selected items, which are maintained locally by different cultural heritage institution.

Challenge: to provide a service to link various cultural heritage providers, such as libraries, museums, archives and so on, by aggregating metadata from them and providing a unified search point to various object collections using that metadata.

3.2.1 Interoperability challenges in National Libraries (from the Koninklijke Bibliotheek National Library of the Netherlands experience)

Challenge 1: Interoperability of objects between (two ways):

- publisher and KB LTP repository (e-Depot)
- between the KB repository and another repository (for example for duplication or back up)
- between the KB repository and a "services" provider

1. **Current situation**:
    a. Very Bad
    b. Bad
    c. **Fair***
    d. Good

    e. Very Good

Short justification: one of the main challenges is the importance of “provenance” information, which informs a library whether the received object is the original objects with an original identification, in order to verify that it is the “authentic” object. Digital objects can easily be replicated however and a “Intellectual Entity” can have different manifestations. A Digital Object Identifier (DOI) will solve some problems, but especially in more complicated cases, this might not be enough. Other challenges are:

- there is not a uniform approach towards adding DOI’s
- although many organizations support the OAIS model, their Archival Information Packages (and Submission Information Packages) differ, this will always lead to conversions in case of exchange of information. There currently are no standard exchange Submission Information Packages for say STM-publications, although many of the publishers use the same standard (NLM)
- versioning of publications, especially in the case of digital objects from publishers.
- finding a “shared language” to streamline internationally used and locally used standards
- the rapidly changing environment in which libraries operate.

Challenge 2: Interoperability of metadata between (one way)
- KB and Europeana
- KB and Clarin

1. **Current situation**:
    a. Very Bad
    b. Bad
    c. Fair*
    d. Good
    e. Very Good

Short justification: both Europeana and Clarin each have developed a normalised format to deal with the Dublin Core (extended) metadata that are in use in the participating libraries. Clarin created a registry to administer the deviations of Dublin Core. Although this solution is a clearly a practical approach of interoperability, one could wonder whether this is a sustainable solution. Versioning is seen here as a special challenge.

2. **Importance** (i.e. impact):
    a. Low
    b. Medium
    c. High*

Challenge 3: Interoperability of objects and metadata
- Researchers and KB research data resources (one way or two ways)

1. **Current situation**:
    a. Very Bad
    b. Bad

c. **Fair***
d. Good
e. Very Good

Short justification: although different initiatives are underway, currently the referencing from publications to related datasets, that were not originally part of the publication and might be preserved in another organization or even another country is still a challenge. As the KB-NL does not preserve datasets, a framework need to be developed to relate KB publications to datasets that are held in another organisation, often, but not always, in the Netherlands.

2. **Importance** (i.e. impact):
   a. Low
   b. Medium
   c. High*

Challenge 4: Interoperability between KB services and a service provider (for example Impact)

1. **Current situation**:
   a. Very Bad
   b. Bad*
   c. Fair
   d. Good
   e. Very Good

Short justification: at his moment there are not yet services developed of which the KB –NL makes use. The KB-NL takes part in the European project SCAPE http://www.scape-project.eu/ . The aim of SCAPE is to develop a framework where large scale preservation actions can be performed. To make this possible, software for both the sender as well as the receiver will be developed. It might well be the case that a library will send a part of its collection to a certain central service provider to have a preservation action performed. At the moment the KB-NL has no intentions yet.

2. **Importance** (i.e. impact):
   a. Low
   b. Medium*
   c. High

Challenge 5: Interoperability between authors

1. **Current situation**:
   a. Very Bad
   b. Bad
   c. Fair*
   d. Good
   e. Very Good

Short justification: the KB-NL, makes use of a local (national) approach by adding a DAI (Digital Author Identifier) to the author names in the shared National Catalogue, especially developed for research publications. This is also related to the VIAF, Virtual International Authority File (currently in a testing phase for the KB-NL).

2. **Importance** (i.e. impact):
   a. Low
   b. Medium
   c. High*

Short justification:

3. **Level of difficulty** (to achieve):
   a. Easy
   b. Fair*
   c. High

3.2.2 Generalized challenges about Semantic Interoperability

1. **To provide mapping between vocabularies, thesauri and categorization systems to facilitate browsing and searching in several library catalogues in parallel with the keywords from any of the used thesauri.**
2. **Aggregating diverse data sources and performing vocabulary alignment to a common ontology in order to facilitate searching and finding structured results also across multiple languages.**
3. **To provide metadata mapping between domain-specific metadata models used by different sources.**
4. **To interlink metadata relevant for digital preservation actions (e.g. metadata about digital objects, their formats, versions, events and agents involved in the events)**
5. **To aggregate metadata from different data providers and provide a common way to search for their content using these metadata.**
6. **To create semantic links between heterogeneous materials from different sources including web resources.**
7. **To provide identification mechanisms for accessing provenance (metadata) of digital objects and intellectual entities.**
8. **To develop a common standard for exchange information between institutions adopting different archival systems.**
9. **To define a framework to relate library publications to datasets that are held by other institutions.**

3.3 SEMANTIC INTEROPERABILITY IN THE EO DOMAIN

Earth Observation resources, products and services are useful for a whole lot of different applications, ranging from security to climate monitoring and from research to operational activities. Each of these application fields has its very specific knowledge and expertise, which does not necessarily include any understanding of the Earth Observation domain. A typical situation is the one described by the following scenario, describing the situation the domain experts, interested in using EO data to support their activities, have to cope with. One way to help them is to set up web services that allow easily identifying and getting access to the EO resources that best fit their needs. In particular, permitting the discovery through controlled terminology the application domain experts are familiar with, and EO resource access through centralised and/or federated catalogues (Heterogeneous Mission Accessibility Cookbook, 2012).

Semantic access to Earth Science Resources

Scenario 9:

Experts of heterogeneous application domains (e.g.: earthquake, fire, flood, etc.) frequently have to cope with a large number of issues for obtaining (discovery and get access) the EO resources they need for their scientific and operational activities. Firstly, they encounter some difficulties in identifying useful resources, as they are typically described with very specific Earth Observation details (e.g. sensor type, resolution, technical characteristics), and hence, they have problem in getting access to them, as data repositories are provided through several catalogues, scattered here and there, having different and frequently non compatible interfaces.

Challenge:

One way to help Application Domain experts in getting access to needed EO resources, might be to set up an interoperable and pluggable architecture, permitting to:

- Discovery data via controlled vocabulary, which would permit the user to search resources through familiar terminology;
- Provide direct access to the needed resource, independently where the resources are physically hosted (e.g.: federation of smaller and remote catalogues)

The environment shall provide at least:

- Ontology mediator capability, permitting to reuse and merge ontologies either complementary to each other, or partially/totally redundant;
- Harvesting capability, permitting to seamlessly collect metadata information from remote and scattered catalogue;

1. **Current situation**:
    a. Very Bad
    b. Bad*
    c. Fair
    d. Good
    e. Very Good

Short justification: the operational use of semantic, within the EO data discovery and exploitation, is currently not mature enough, as so far, only demonstrators, prototypes and / or small and customised solutions are proposed to the EO community. In spite of the relevant effort spent in the last years, the lack of a common understanding of the problem, common requirements and standardised technologies has considerably delayed its adoption in the EO domain context.

2. **Importance** (i.e. impact):
    a. Low
    b. Medium
    c. High*

Short justification: The relevance of the challenge is extremely high. Think about the amount of data collected into the GMES repository, or Sentinel-1/2, Cosmo-SkyMed, and many other upcoming missions. The ontology would permit the user, first to discovery, and then to get access to the data, much more easily and quickly with regard to conventional catalogues. Furthermore, ontology mediation techniques would permit a user, belonging to a scientific community A, to get access to the data made available from the scientific community B, still using its own terminology.

3. **Level of difficulty** (to achieve):
    a. Easy
    b. Fair
    c. High*

Short justification: a step in the right direction is being done within the Heterogeneous Multi-mission Accessibility (HMA) initiative. HMA initiative is mainly based on the involvement of key stakeholders, such as national space agencies, satellite or mission owners and operators, and industry for successfully addressing this challenge. Participants in the HMA Initiative have also been actively engaged in the OGC, contributing to the development of OGC standards and successfully aligning European Earth observation systems with standards from the OGC, and related standards organizations such as those from the International Organisation for Standardisation (ISO).

Anyway, with the aim of finalising the work done up until now, it will be necessary to promote and adopt as much as possible these standardisation outcomes. This will permit to migrate from a current prototyping / demonstration phase, into an operational one.

Barriers preventing the re-use of existing ontologies

Different sources can describe their resources with heterogeneous semantic terms, which can cause confusion (similar but not identical meaning), scaling and unit conflicts (use of different reference systems) or naming conflicts (e.g.: due to homonyms and synonyms). In order to exploit at maximum the benefits of data from all sources, new methodologies must be investigated and implemented to ease relevant data identification, through semantic terms meaningful for different user domains. The preferred approach is through an ontology-based system, relying on common terminology, and permitting access either to computers or human beings.

For instance if applied to ground segments, an ontology based system can enlarge the use of the outcome of EO missions by easing the semantic identification (also from non-EO domains) of relevant / useful EO resources. It is a priority to understand that even for a single and well-defined application domain, such as EO ground segment, many ontologies could be available. This is because different persons and organisation, with different opinions, languages, technology and point of view of a same topic can be involved within the ontology development. As a consequence, such ontologies could contain conflicts and overlaps, and different ontologies could describe the same domain, or parts of, in a different way. The issue can be only partially mitigated with the use of Ontology Mediation techniques.

Ontology mediation is a broad field of research which is concerned with determining and overcoming differences between ontologies in order to allow the reuse of such ontologies, and the data annotated using these ontologies, throughout different heterogeneous applications. It can be subdivided into three areas:

- ontology mapping, which is mostly concerned with the representation of correspondences between ontologies;
- ontology alignment, which is concerned with the (semi-)automatic discovery of correspondences between ontologies;
- ontology merging, which is concerned with creating a single new ontology, based on a number of source ontologies.

Going through the topic more in detail, the resources discovery, within the Earth Observation domain, is typically performed using catalogues containing information about EO pertinent services, image collections, products, and sensors. Catalogues typically store metadata describing these resources and allow for the catalogue user to search for resources of interest using a set of criteria. Depending on the type of the resource and the user community (e.g.: GMES, INSPIRE, etc.), different types of catalogues storing different kinds of metadata are used (e.g: ISO, INSPIRE, CSW-ebRIM, etc.). In order to enhance the classical discovery mechanism, ontologies can be used to provide keywords. The catalogue user then uses these keywords as queryables and the catalogue returns the metadata records which contain these keywords in their metadata records. This scenario, however, assumes the use of Ontologies that are compatible with the keywords used in the resource metadata. These ontologies define terms that are not necessarily familiar to the catalogue user, hence a data mediation is required when the semantic content of data required by systems is the same, but the syntactic representation is different. For example, within the SMAAD project an ontology mediator shall be set up with the aim of semantically mapping the INSPIRE GEMET Ontology and EO Ontology (see RARE projects), using this mapping, a query using terms from GEMET ontology permits to discover GSDCA metadata, and a query using GSDCA term permits to discover an INSPIRE metadata.

Once the ontology has been designed, collecting terms and reciprocal relationships, the following steps is the choice of a suitable semantic representation language. This is necessary to formalize the definition of semantic entities and their relationships in formats that are machine-readable and –processable (see section 4.4 for more detail about technologies). The principal set of representation languages are all recommendations of the World Wide Web Consortium (W3C), such as RDF / RDFS, OWL / OWL2, XMT – XML for Topic Maps and SKOS. This is not a bias towards the W3C, but is due to the fact that these specifications, all defined in the scope of the "Semantic Web" vision developed at the W3C, have overridden other projects over the past years, and become the de-facto set of standards upon which any further standards are now built. The choice of the language should not be underestimated, it should take care about the "level of interoperability" of the technology adopted for the thesaurus / ontology / dictionary implementation, and the compatibility with web client applications (e.g.: an available web client, served by an OWL based Web Service, cannot ensure a direct reuse of SKOS thesaurus, unless of ad-hoc refinement)

Once the thesaurus has been developed, it is necessary to link it to the resource repository, e.g. catalogue, following a "semantic annotation" procedure. This issue consists in annotating the resource metadata with pre-defined semantic keywords, permitting to set up a relation between the resource and the ontology entity (e.g. linking an ISO19115 metadata file with an ISO Topic defined in a SKOS thesaurus). The type of that relation has to be clearly defined and agreed upon, so that these annotations are interpretable, indexable and queriable later on.

Is digital preservation addressed?

The huge amount of available EO data, make the activities of coordination and harmonisation of Long Term Data Preservation (LTDP) ESA Programme almost mandatory, in particular if we consider the upcoming Sentinel missions. At European level the LTDP aims to:

- Preserve the European, and Canadian, EO space data sets for an unlimited time-span;
- Ensure and facilitate the accessibility and usability of the preserved data sets respecting the individual entities applicable data policies;
- Through the adoption of a cooperative and harmonized collective approach among the data owners (LTDP Framework) based on the application of European LTDP Common Guidelines and sustained through cooperative (multi-source) long term funding schemes;
- Ensure, to the maximum extent, the coherency with the preservation of other non-space based environmental data and international policies.

The approach proposed by the LTDP aims at the progressive application of the European LTDP Common Guidelines but also at cooperation of the archive owners in several areas for a progressive development and implementation of technology, methodology, standardization, operational solutions and data exploitation methodologies as key aspects for the set-up of the framework.

For the time being, none of the described “semantic projects” directly addresses the LTDP objectives, but on the other hand, LTPD architecture will be also based on catalogue, with standardised interfaces and protocols if available. This should make possible, at least theoretically, a partial reuse of current “semantic projects” outcome, as the idea behind the described semantic access means is the availability of an ontology, independent by the specific data repository, and a its link to standardised catalogue interfaces. So, in principle, there is no particular reason of thinking about an incompatibility between available semantic access means with what will be a standardised LTDP specific catalogue.

3.4 PROVENANCE AND AUTHENTICITY INTEROPERABILITY

Several provenance models have been proposed in a number of domains (e.g. scientific workflow, databases, Semantic Web) with different levels of expressivity, different perspectives (e.g. agent-centered, object-centered or process-centered) and assumptions about the system they are embedded in. The idea that a unique global model for representing provenance information could be adopted by all systems seems an unrealistic option today. Therefore, interoperability between current models is a crucial issue to allow that provenance information can be exchanged among heterogeneous systems, which internally use different provenance models.

This section reports three scenarios illustrating three interoperability challenges (integration, querying and searching) concerning provenance information.

Exchange and Aggregation of Provenance Information:

Scenario 10

A sensor e.g. at a satellite, makes some measurements. The measurements are then transferred to a ground station. The data are then processed by a group of researchers, say group A, to produce an image, say img1. The image is then processed by group B to produce a second image, say img2.

To produce the complete provenance of the img2 (which may be important for assessing the credibility/authenticity of img2) we have to aggregate the provenance information of each data object and link them appropriately. This aggregation requires having a common model for representing provenance or mappings between the adopted models.

Challenge

Ability to exchange and aggregate provenance information of various processing tasks or transfer/archiving events.

1. **Current situation**:
   a. Very Bad
   b. Bad*
   c. Fair
   d. Good
   e. Very Good

Short justification: Although there are several models for modelling and recording provenance, and various mappings between these models have been defined (as described in D24.1 of APARSEN WP24), their adoption by the various organizations has not progressed as it could. Probably the prominent importance of provenance information for e-Science has not been realized.

2. **Importance** (i.e. impact):
   a. Low
   b. Medium
   c. High*

Short justification: Provenance information is of prominent importance for e-Science (e.g. for checking and validating results, for reproducing them, etc).

3. **Level of difficulty** (to achieve):
   a. Easy
   b. Fair*
   c. High

Short justification: The technical solutions to the challenge are not very difficult to implement. There is extensive literature on how to exploit mappings for rewriting queries or information and there are several tools that can aid such tasks (more information is provided in the internal deliverable ID2401 of APARSEN WP24). Furthermore, a method that can aid the ingestion and management of provenance information (specifically a rule-based method for propagating provenance information) has been elaborated in the context of APARSEN, see (Strubulis, Tzitzikas, Doerr, & Flouri, 2012).

Query of Provenance Information

Scenario 11:

A user is viewing using his web browser an image about the ozone hole. He wants to be sure about the credibility of the image. Ideally he would like to have a button, say "get provenance", that would initiate the process of getting provenance information about this image. On clicking, a number of basic queries (like those specified at (Theodoridou, Tzitzikas, Doerr, & Marketa, Modeling and Querying Provenance by Extending CIDOC CRM, 2010) are applied over the provenance information of the image. This information could be embedded as metadata in the image format itself, or fetched from external sources. The obtained query results could contain references to other digital objects for which the user should be able to click again the "get provenance" button, and so on. Ideally, it should be possible to reach the raw data that were used for deriving the image that depicts the ozone hole, i.e. those obtained by the sensors of the satellites, and get information about these sensors too. For instance it should be possible to get information about the context (e.g. information about the orbit of the satellite that hosts the sensor).

Challenge

Ability to apply some basic queries over the provenance record of any digital object. Services that can fetch and integrate the required provenance information from heterogeneous and distributed sources.

1. **Current situation**:
   a. Very Bad
   b. Bad*
   c. Fair
   d. Good
   e. Very Good

Short justification:

The provenance information of some digital objects, even if it exists, it is in many cases not public. In the ideal case such information should be public, and dedicated services for provenance information should become the common practice.

2. **Importance** (i.e. impact):
   a. Low
   b. Medium
   c. High*

Short justification:

The relevance of the challenge has become very high in the last years with the growing amount of digital material. It is more and more harder to distinguish the credible material from the fake material. This problem will be alleviated if provenance information becomes public and in a machine processable way.

3. **Level of difficulty** (to achieve):
   a. Easy
   b. Fair*
   c. High

Short justification: a technical solution to the challenge is not very difficult to implement but it is worth to say that the success of the solution is far to be a mere technical problem, concerning, economical, political and social dimensions. The Linked Data initiative has produced practices and tools that can aid the process of publishing provenance data.

Finding information about a resource's authenticity and availability (from D22.1)

Scenario 12

Paul is a researcher at the University of Amsterdam. His current research interest is the behaviour of politicians before elections and after they are elected. To investigate this behaviour he uses as input political programs (what did they promise before the elections) and minutes from the parliament (what did they actually do after being elected). Both resources are accessible on the Internet. Paul wants to use these resources. For Paul it is important that the resources are authentic and that they will be available not only today, but also in 10 years time.

Based on these two resources he publishes his results.

Paul wants to refer in his publication to these resources for reasons of verification and validation for his peer reviewers.

For Paul it is important to know:

- Are the resources authentic?
- Are the resources permanent accessible?
- Where can a resource be retrieved?
- How to refer to the resources?

A trusted persistent identifier infrastructure could provide Paul with answers to these questions. Paul can go to a global resolution service and retrieve information about the organization that authorized the PID. When the authorizing organization is part of the trusted persistent identifier infrastructure then Paul knows it is an organization he can trust. Paul can also retrieve the information of the organization regarding their policy on permanent access. The trusted persistent identifier infrastructure supplies Paul with good practices on how to reference the resources.

Challenge: to certify, trough a trusted infrastructure, the authenticity and accessibility across time of Web resources distributed across systems.

1. **Current situation**:
    a. Very Bad
    b. **Bad***
    c. Fair
    d. Good
    e. Very Good

Short justification: The complexity of the preservation function in the digital area requires the development of specific tools able to ensure that the main elements and procedures relevant for the quality of the preservation are maintained, and the authenticity of the preserved information objects can be presumed. Today, some solutions and models have been proposed to address the authenticity challenge in the domain of digital preservation. CASPAR for example has identified the need for an Authenticity Management Tool with the capacity of monitoring and managing protocols and procedures across the custody chain in order to deliver the benefits of authenticity into information systems, from the creation to the preservation phase.

2. **Importance** (i.e. impact):
    a. Low
    b. Medium
    c. High*

Short justification: maintaining the authenticity (trustworthiness) and access of the preserved digital objects for the long term, during which technologies, formats, hardware and software are very likely to change, is of great importance, since users must be confident that the objects in the changes environment are authentic.

3. **Level of difficulty** (to achieve):
    a. Easy
    b. Fair*
    c. High

Short justification: finding a solution for authenticity management is complicated by a number of aspects:

- Authenticity fundamental requirements must be clearly identified in order to avoid at the same time overload and lack of information.
- Authenticity methodology and concepts are cross-domain but their deployment is strongly dependent on specific environments.
- Integration of concepts from different ontologies may be difficult due to their overlapping.

3.4.1 Generalized challenges about Provenance

1. **To develop a common model for representing provenance information or a mapping solution between different models to aggregate provenance information from different sources.**
2. **To provide query and retrieving systems and user interfaces to give access to heterogeneous and distributed provenance information.**
3. **To develop a trusted PIDs infrastructure which guarantees access to authentic digital objects and related provenance information.**

3.5 DIGITAL PROPERTY AND RIGHTS MANAGEMENT INTEROPERABILITY

Rights Management

The protection of intellectual property rights and controlled access to resources is a fundamental issue for many cultural heritage institutions (e.g. libraries, archives, research institutes) especially in the digital era where digital data are highly shared and transmitted across system boundaries.

Therefore, approaches for digital property and rights policy interoperability represent a big challenge for preserving the accessibility and reusability of digital content in the long term. In this section we describe a scenario, which can be considered a metadata interoperability scenario, where rights interoperability is involved.

Scenario 13[21]

Anoushka is carrying out research on a topic which is developing rapidly, so she needs to access a range of online journals, databases, and other information resources. Her institutional library uses a system which matches her user profile with multiple access conditions imposed by journal vendors, so she can clearly see what sources of information are available to her. Anoushka works from home for one or two days per week, and finds that some vendors impose additional access restrictions for users outside of the institution’s IP networks. She intends to undertake a three-month study tour of other institutions engaged in similar research, and is uncertain if she will be able to access the information she needs from their systems. She does not want to spend too much time registering her profile in each institution’s library, or rely on a resource which becomes unavailable as she moves from library to library. She is concerned that she may spend time searching, identifying, and selecting a resource which she then finds she cannot access.

Challenge: a standard way to expose rights expressions with metadata is needed. However several limitations and problems can due to:

- Geographic restrictions may be not explicitly stated
- Library subscriptions change frequently
- Copyright status may not be clear

4 INTEROPERABILITY SERVICES, STANDARD MODELS, FORMATS AND VIRTUALISATION MODELS AND INTERFACES

The brief introduction to the use of the interoperability term reported in Section 2 shows that interoperability is a very complex concept, which includes several aspects and nuances to be considered of which the most usually cited one of technology is normally the most straightforward to solve. Very often these aspects of interoperability are presented as separate but dependant layers: the higher level of interoperability is possible only if the lower level is ensured (see for example (Paskin, 2006) about identifier interoperability). In this work we propose to use a slightly different lens to examine the interoperability issues from a digital preservation point of view.

4.1 FOCUS AND METHODOLOGY

Since our goal was to investigate the interoperability issues encountered by the APARSEN partners as part of daily activities and gather the conceptual models, services and standards used by them to address these issues, we aimed at identifying concrete interoperability issues, needs and related solutions (i.e. models, standards, frameworks, services) adopted by the partners in relation to the key digital preservation areas investigated within the APARSEN project.

[21] http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Digital_resources_with_access_restrictions

The final aim was to describe which are the critical interoperability aspects pertaining in a certain area of digital preservation, which main layers of interoperability are mainly involved, which are the interoperability objects that are implicated and finally which concrete solutions (e.g. models, standards) have been adopted to address these issues. The result of the analysis led to define a sort of matrix, which combines different layers of interoperability (e.g. syntactic, semantic, organizational) with the areas of digital preservation (e.g. persistent identifiers, metadata, provenance) and the related interoperability objects and models, providing an interoperability conceptual framework for digital preservation that can be used as a starting point to facilitate practical interoperability solutions and design concrete interoperability services for long-term preservation.

Since several specific interoperability issues and solutions have been already studied in (or are the current focus of) other WPs of the project and are experienced by many partners in their daily work activities, we started to exploit this internal knowledge by asking the APARSEN partners to provide some preliminary information on this topic.

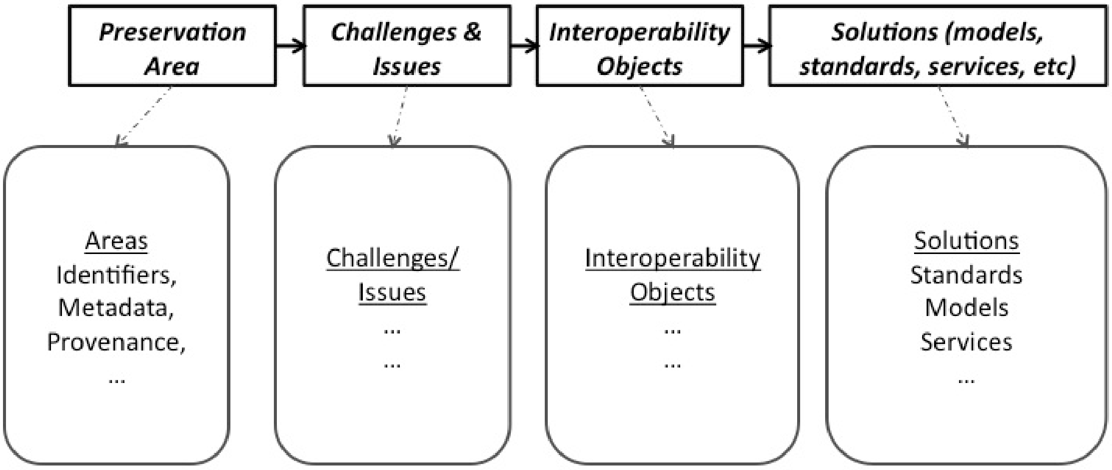
To this purpose, we proposed to organize the provided information on the basis of a common framework that aims to characterize the problem facets as well as the existing and forthcoming solutions and models. In this way the specific challenges of interoperability within a specific area could be directly linked with the current available solutions. Partners were asked to provide relevant information (according to their knowledge or experience) about the interoperability issues within a given sector of digital preservation and list related models, services and standards that are in use within their organization to address the identified challenges. In particular we asked to provide information by using the following four categories:

1. **Digital preservation area:** indicates the area of digital preservation where interoperability takes place. Looking at the organization of the APARSEN project, examples of digital preservation areas could be preservation services, persistent identifiers and citability, storage solutions, authenticity and provenance, annotation, reputation and data quality, scalability, digital rights and access management, 3rd party certification of repository.

2. **Interoperability issue/challenge:** a problem of interoperability which hinders a certain task or process in an interoperability context.

3. **Interoperability objects:** are the entities that actually need to be processed in interoperability scenarios. They can include for example the full content of digital resources or mere representations of such resources (i.e. metadata, identifiers).

4. **Adopted solutions/ models/ standards:** are those approaches, which are adopted to address specific interoperability issues/challenges at different levels.



Finally partners were asked to link relevant resources (e.g. deliverables, either internal or external to the APARSEN project, papers, project/initiatives Websites and so on) on interoperability models, standards and services in digital preservation.

As final remark, it is important to notice that the focus of this section is specifically on interoperability solutions for digital preservation. We refer to other project deliverables (in particular the D13.1) for a more extensive overview about standards dedicated to preservation activities in general.

4.2 INTEROPERABILITY SOLUTIONS FOR DIGITAL PRESERVATION

The following figure provides a mind map that summarizes the contents of the material that follows. The solutions have been clustered around eight categories identified by colours (maintained in the matrix below). As example, the Persistent Identifiers node in the graph has been exploded in Figure 13 to show some additional details. This gives an idea of how the collected information can be organized into a search tool, which allows to browse the information space. Potential benefits of this and other tools built on the results of the present work will be explained in Section 6.

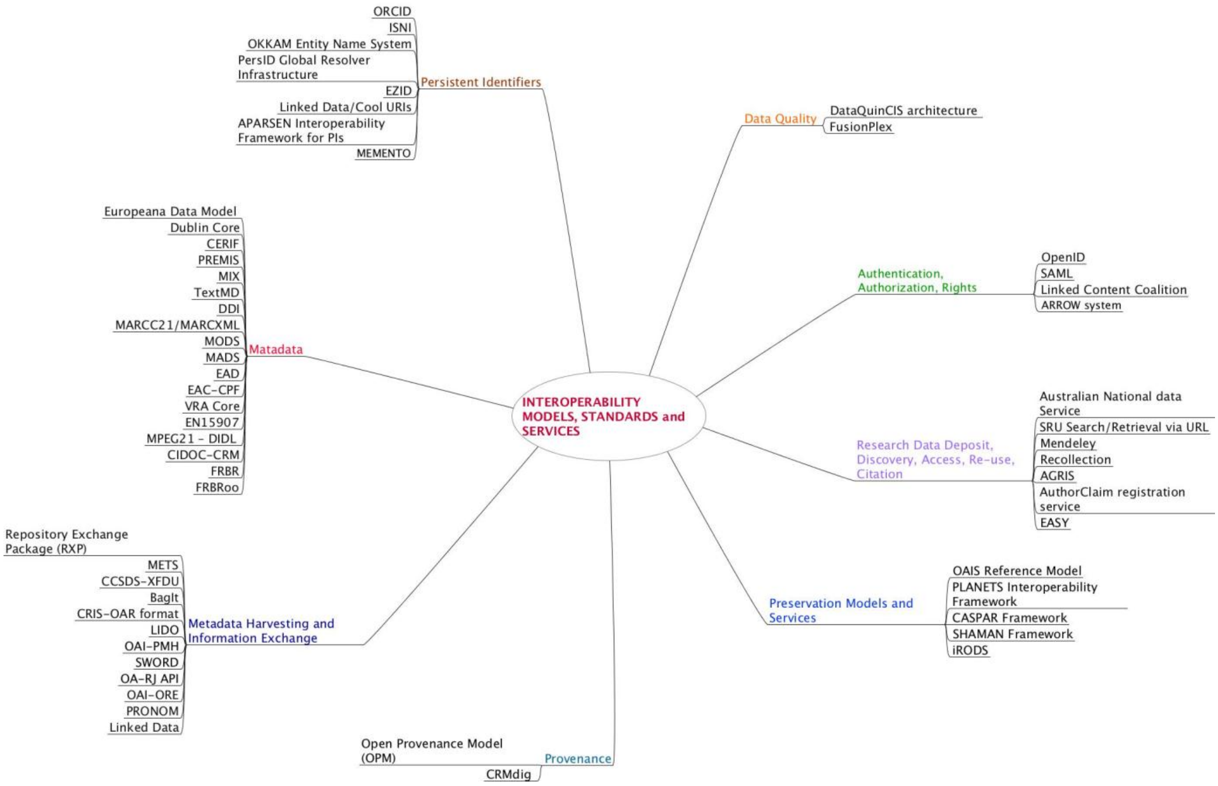


Figure 12: Interoperability Solutions

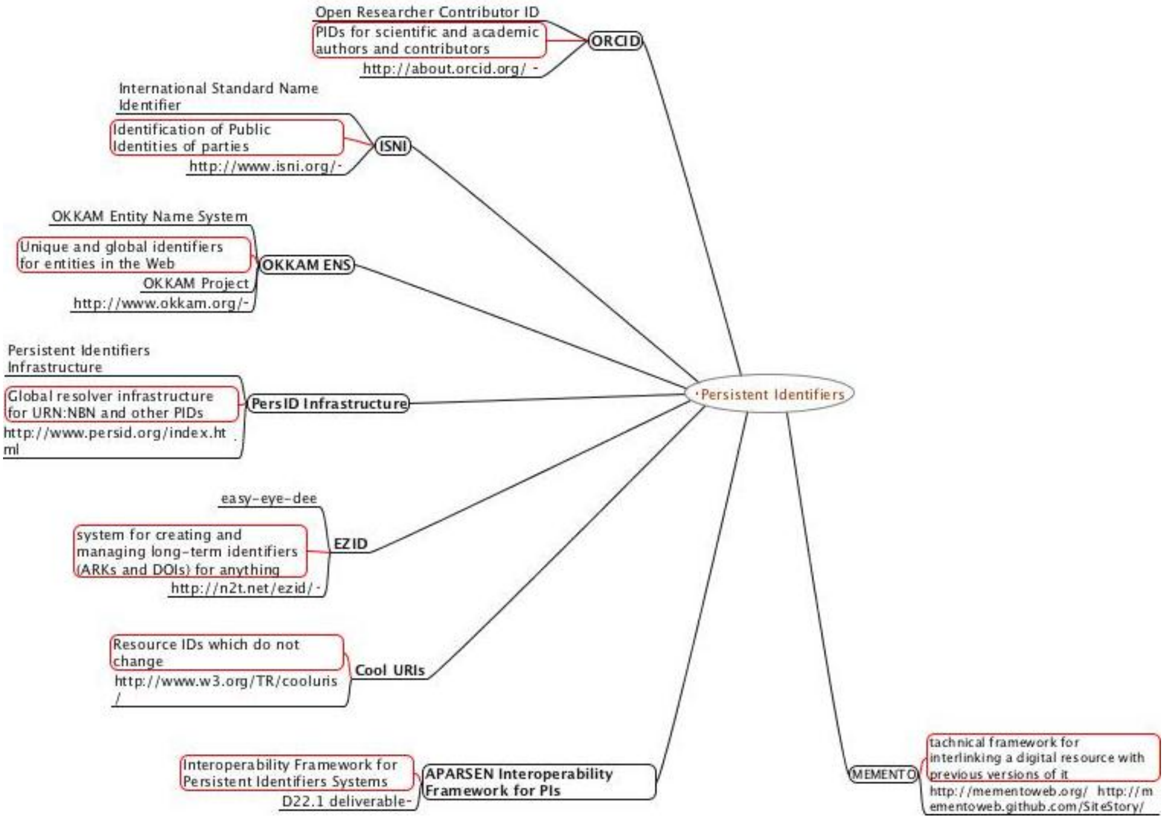


Figure 13: Persistent Identifier Interoperability Solutions

The following table reports on the interoperability solutions, i.e. models, standards, services, virtualization interfaces reported by the APARSEN participants by linking these solutions to the Digital Preservation Area, the related interoperability challenge, the Interoperability objects and the layers of interoperability involved.

The identified Digital Preservation Areas and not independent, but connected in various ways and therefore a solution that has been associated to a certain area (e.g. PIDs) could be indirectly be associated to another area (e.g. Provenance) due to the dependencies between them. For example a solution for enabling the interoperability between PIDs systems has a crucial impact on (and therefore can be connected to) provenance integrating solutions. A diagram that shows some basic relations follows.

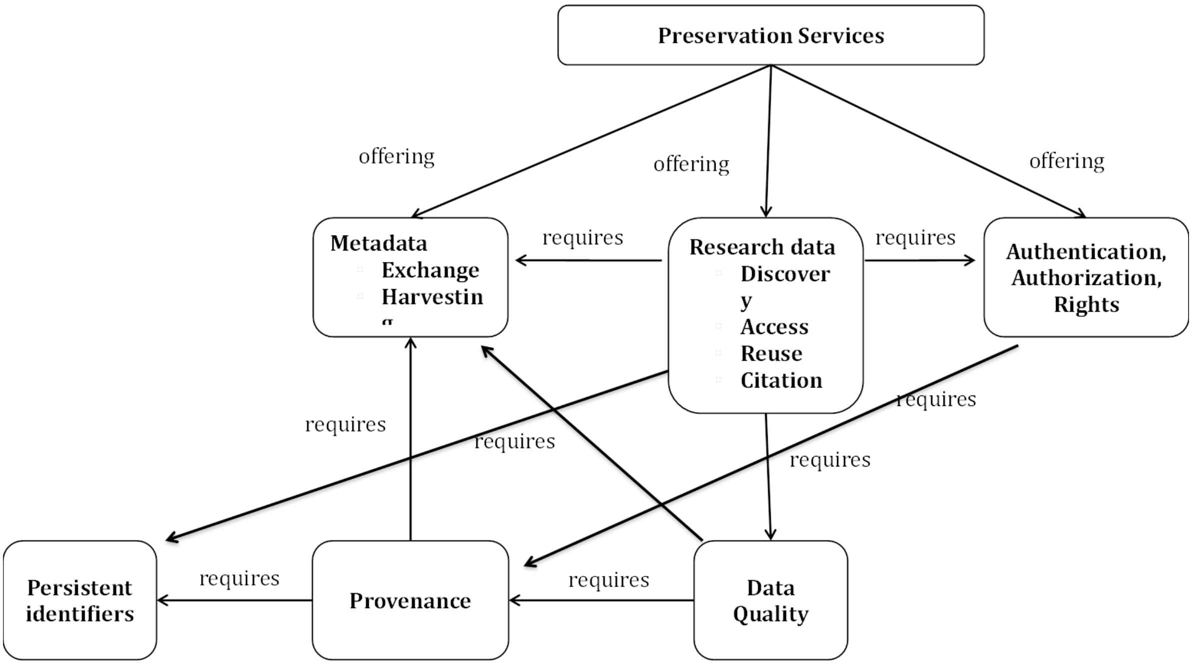


Figure 14: Interdependencies between DP areas

As final remark, it should be noted that some of the interoperability solutions reported below have been developed by projects and initiatives described in more details in Section 2.4. Therefore, the reader can find a general description of the initiative (e.g. International Working Group on FBRB/CIDOC CRM Harmonization) in Section 2.4 and specific information about the related interoperability solution (e.g. FRBoo) in this section. An example is provided in Figure 15 where a number of relationships between concepts of the interoperability landscape are shown, e.g. the Initiative concept is related to the Solution concept, the DP community concept is related to the Solution concept, the Solution concept is related to the Layer of Interoperability.

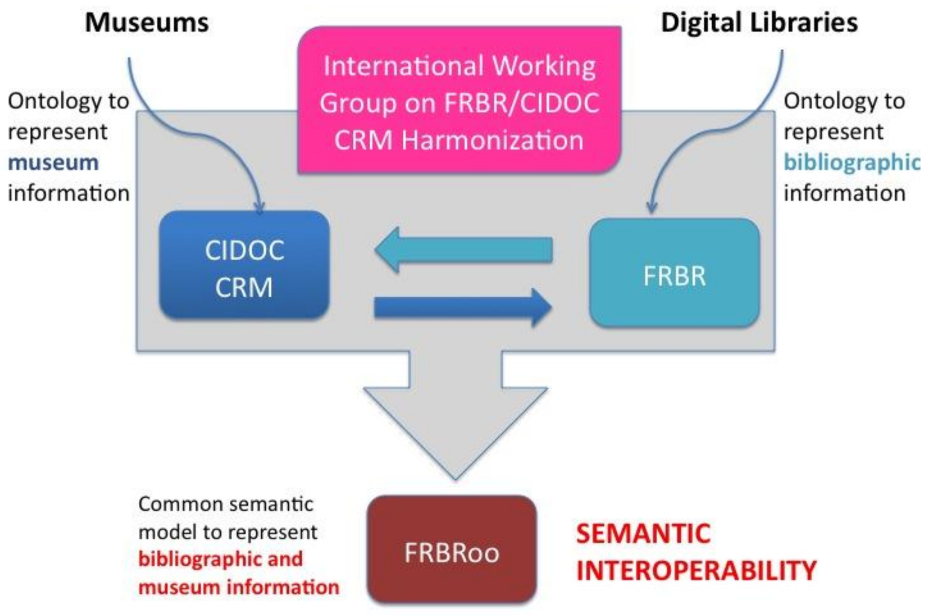


Figure 15: An example of the link between Initiatives and Services described in the document

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
DP Area: PERSISTENT IDENTIFIERS			
Interoperability between persistent IDs for researchers	Persistent IDs for researchers	Syntactic/Semantic	**ORCID** registry: a registry of persistent unique identifiers for individual researchers and an open and transparent linking mechanism between ORCID, other ID schemes, and research objects such as publications, grants, and patents. http://about.orcid.org/
Interoperability between proprietary right holder identification systems (e.g. IPI) and research discovery tools (e.g. VIAF)	Identifiers for public identities of parties	Syntactic/Semantic	**ISNI**: The International Standard Name Identifier is an ISO Standard (ISO 27729) which allows the identification of Public Identities of parties: that is, the identities used publicly by parties involved throughout the media content industries in the creation, production, management, and content distribution chains. It also provides a single identifying code for a party involved in multiple creative genres (music, cinema, visual arts, literature, etc.). The ISNI has been designed as a bridge identifier that connects proprietary author identification systems across multiple public databases. ISNI provides an open layer above proprietary Party identification systems to link all the manifestations of a party in different systems, making it possible for industry partners to exchange party information without disclosing confidential information. http://www.isni.org/
Interoperability for identifiers for entities	Identifiers for any entity named on the Web	Semantic/Syntactic	**OKKAM Entity Name System (ENS):** it is a scalable infrastructure for managing and reusing unique and global identifiers for entities. By linking the OKKAM ID to alternative IDs for the same entity, the ENS provides an interoperability layer for PIDs. http://www.okkam.org/
Interoperability for PIDs for scholarly and cultural information.	Identifiers for scholarly and cultural digital contents.	Semantic/Syntactic	**PersID global resolver infrastructure:** a meta-resolver to provide a unified interface to a network of URN national local resolvers. http://www.persid.org/index.html
Interoperability	ISO TC46/SC9	Syntactic	**Use cases for interoperability:**

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
across the family of ISO TC46/SC9 identifiers	identifiers	Technical	http://www.dlib.org/dlib/april06/paskin/04paskin.html
Assigning and managing unique long-term identifiers	Unique identifiers for anything but special focus on dataset	Syntactic/Semantic	**EZID:** is a service developed and maintained by the UC Curation Center (UC3), a group within the California Digital Library focused on developing systems and tools that enable the curation and preservation of digital scholarship. EZID is a system to create and manage unique, long-term identifiers. It allows to create identifiers for anything choosing from a variety of PIDs (e.g. ARK, DOI), store citation metadata for these identifiers and manage these identifiers for example by updating URL location to make them resolvable for the long term. http://n2t.net/ezid/
To create an interoperability platform for trusted Persistent Identifiers systems	Persistent Identifiers	Syntactic/Semantic/ Organizational	**APARSEN Interoperability Framework for Persistent Identifiers systems:** a general framework, which enables interoperability between PIDs, and a demonstrator (with basic services) have been developed within the WP22. http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework
Identification of and Access to versions of Web resources	Versions of Web resources	Technological Syntactic	**Memento** is a HTTP-based technical framework that bridges the present and past Web by interlinking resources with resources that encapsulate their past. Through the URI of a Web resource, memento allows to see the version of that resource as it existed at some date in the past. Based on this framework the **SiteStory transactional archiving solution** has been developed to enable the creation of dynamic Web archives, which are representative of a server's entire story, by archiving all the versions of the server's resources being requested across time.

Interoperability issue/challenge	**Interoperability objects**	**Interoperability level**	**Adopted solution/ model/ standard /service**
			http://mementoweb.org/ http://mementoweb.github.com/SiteStory/
	DP Area: PROVENANCE		
Interoperability and exchange of provenance metadata	Provenance metadata	Semantic	The **Open Provenance Model (OPM)** is a community-driven model for provenance, which originates from the Provenance Challenge series, allowing provenance to be exchanged between systems. The model provides a common semantics, based on annotated causality graphs, which allows for expressing all the causes of an item. The model aims at supporting a digital representation of provenance for any "thing", whether produced by computer systems or not (i.e physical, digital or abstract entities). http://openprovenance.org/
Interoperability between provenance metadata	Provenance metadata of digital objects for e-science	Semantic	**CRMdig** is a digital provenance model (an extension of the CIDOC CRM ontology) that has been created with the specific purpose of sharing provenance metadata for e-science digital objects. http://www.usenix.org/events/tapp11/tech/final_files/Doerr.pdf
Interoperability between provenance metadata	Provenance metadata	Semantic	Mapping between provenance models (wp 24: mapping between OPM and CRMdig). Details at the APARSEN internal deliverable ID2401 (the external deliverable D24.1 summarized the results). See in particular section 7 of D24.1: http://aparsen.digitalpreservation.eu/pub/Main/ApanWp24/APARSEN-REP-D24_1-01-2_2.docx
	DP Area: DATA QUALITY		
Defining Interoperability framework for CISs (Co-operative Information System)	Data quality metadata	Semantic	**DaQuinCIS architecture** is a platform for exchanging and improving data quality in CISs through a set of data quality services. Organizations export data and quality data according to a common model referred as Data and Quality Model (D2Q) which includes 1) definitions of constructs to represent data 2) a common set of data quality properties 3) constructs to represent

Interoperability issue/challenge	**Interoperability objects**	**Interoperability level**	**Adopted solution/ model/ standard /service**
			them 4) the association between data and quality data. www.dis.uniroma1.it/~midlab/articoli/SVMMB_IS.pdf
Resolution of data inconsistencies in the integration of heterogeneous information sources	Information features (i.e. metadata) about data quality	Semantic	**Fusionplex** is a system for integrating multiple heterogeneous and autonomous information sources that uses data fusion to resolve factual inconsistencies among the individual sources. To accomplish this, the system relies on source features, which are meta-data on the quality of each information source; for example, the recentness of the data, its accuracy, its availability, or its cost. cs.gmu.edu/~ami/teaching/infs797/current/**fusionplex**.pdf
DP Area: METADATA			
Interoperability between metadata schema and vocabularies	Metadata	Semantic	**Europeana Data Model (EDM)** is a cross-domain Semantic Web based framework for collecting, connecting and enriching metadata from multiple content providers. To this purpose the model includes a well-defined set of elements from two categories: 1. The elements re-used from other namespaces: RDF and RDF Schema; OAI ORE; SKOS; Dublin Core 2. The elements introduced by the EDM http://pro.europeana.eu/edm-documentation
Metadata standard	Metadata	Semantic	**Dublin Core**: is a metadata standard providing a vocabulary for the description of any kind of entity like physical objects, digital objects and compound objects. http://dublincore.org/
Defining a common representation of research	Metadata representing research entities	Semantic	**CERIF:** it is a standard maintained by European Organization for International Research Information to represent research entities and their relationships.

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
entities and their relationships			It is intended to provide an intermediate layer of interoperability for EU research information. http://www.eurocris.org/Index.php?page=featuresCERIF&t=1
Metadata standard for digital preservation	Preservation metadata	Syntactic/Semantic	**PREMIS** (PREservation Metadata Implementation Strategies): International XML-based metadata standard, which supports digital preservation and ensures the long-term usability of the preserved digital objects. PREMIS is supported by several tools and systems related to digital preservation. http://www.loc.gov/standards/premis/
Metadata standard for images and image collections	Metadata for images	Syntactic/Semantic	**MIX**: XML schema for representing the technical metadata of digital still image objects. http://www.loc.gov/standards/mix/
Metadata standard for text	Metadata for text-based objects	Syntactic/Semantic	**TextMD**: XML schema for representing the technical metadata of digital text-based objects. http://www.loc.gov/standards/textMD/
Metadata international standard for describing data from the social and behavioural sciences	Metadata for research data	Syntactic/Semantic	**DDI** (Data Documentation Initiative): XML format for describing data from the social, behavioural, and economic sciences. It accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing and archiving. The specification supports the entire research data life cycle. http://www.ddialliance.org/
Metadata standard for bibliographic information	Bibliographic Metadata	Syntactic/Semantic	**MARC21/MARCXML**: MARC21 is a representation and communication standard designed for the bibliographic and related information. MARCXML is the XML schema for MARC data. http://www.loc.gov/marc/ http://www.loc.gov/standards/marcxml/

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
Metadata standard for bibliographic sets	Bibliographic Metadata	Syntactic/Semantic	**MODS** (Metadata Object Description Schema): MODS is a schema for a bibliographic element set, and it is designed particularly for library applications. http://www.loc.gov/standards/mods/
Metadata standard for agents and events	Metadata for agents, events and terms	Syntactic/Semantic	**MADS** (Metadata Authority Description Schema): XML schema that may be used to provide metadata about agents, events and terms. MADS can be used independently or with MODS schema. http://www.loc.gov/standards/mads/
Metadata standard for describing information about archival collections	Metadata about collections	Syntactic/Semantic	**EAD** (Encoded Archival Description): EAD is an online environment standard for the encoding inventories, indexes and guides, which are created by archival and manuscript repositories for providing information about specific collections. http://www.loc.gov/ead/
Metadata standard for creators of archival material	Metadata forn creators'names	Syntactic/Semantic	**EAC-CPF** (Encoded Archival Context for Corporate Bodies, Persons, and Families): XML schema for names of creators of archival materials and related information. http://eac.staatsbibliothek-berlin.de/
Metadata standard for description of works in the domain of visual culture	Metadata for visual data	Syntactic/Semantic	**VRA Core**: Data standard for the description of works of visual culture including the images that document them. http://www.loc.gov/standards/vracore/
Metadata standard for cinematographic works	Metadata for cinematographic works	Syntactic/Semantic	**EN 15907**: European standard, which defines a metadata set for describing cinematographic works. http://filmstandards.org/fsc/index.php/EN_15907
Standard representation and identification of	Metadata Digital items	Syntactic/Semantic	**MPEG21 – DIDL (Digital Item Declaration Language)** DIDL is a language providing a markup

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
digital objects			representation for the Digital Item Declaration abstract model of MPEG-21 Framework. A digital item in this framework is a structured digital object, including a standard representation, identification and metadata. A DIDL document is a single hierarchical XML tree of digital resources, with metadata nested within the objects to which they refer. The top level element is <DIDL> which must contain exactly one container or item. Metadata concerning the DIDL document as a whole are provided by the <DIDLInfo> element. http://xml.coverpages.org/mpeg21-didl.html
Common semantic framework for representing cultural heritage information	Concepts and relationships used in cultural heritage documentation	Semantic	**CIDOC Conceptual Reference Model (CRM)** is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information by providing a common and extensible semantic framework for describing concepts and relationships used in cultural heritage documentation. The CIDOC CRM can be conceived as an interoperability framework which provides the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives. http://www.cidoc-crm.org/index.html
Conceptual model for bibliographic information	Concepts and relationships in the bibliographic universe	Semantic	**FRBR (Functional Requirements for Bibliographic Records)** is an entity-relationship model developed by IFLA study group intended to be independent of any cataloguing code of implementation. http://www.ifla.org/publications/functional-requirements-for-bibliographic-

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
			records
Integration and interchange of bibliographic and museum information through harmonization of reference models used by the two communities	Concepts and relationships used in libraries and museums	Semantic	**FRBoo** is a formal ontology, developed as joint effort of CIDOC Conceptual Reference Model and Functional Requirements for Bibliographic Records International working group (see International Working Group on FRBR/CIDOC CRM Harmonisation in Section 2.4 for more details), intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information. The model provides a solution to the problem of semantic interoperability between the documentation structures used for library and museum information. http://www.cidoc-crm.org/frbr_inro.html
DP Area: METADATA HARVESTING and INFORMATION EXCHANGE			
Data and metadata exchange	Metadata Data Archival information packages	Syntactic/Semantic	**Repository Exchange Package (RXP),** developed within the TIPR Towards Interoperable Preservation Repositories project http://ddp.nist.gov/workshop/papers/03_08_Caplan_TIPR.pdf The TIPIR project, which is reaching the midpoint of its second year, has drafted, implemented, and tested a specification for a Repository Exchange Package (RXP), a hierarchical packaging format designed to facilitate the exchange of Archival Information Packages (AIPs) between digital repositories. The RXP encodes structural and preservation metadata using METS and PREMIS, two widely used schema in the cultural heritage community. It is agnostic to the application software used by the sending or receiving repositories or the number of representations included in any AIP.

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
Metadata standard for encoding and exchanging information about digital library objects	Metadata representing digital library objects	Syntactic/Semantic	**METS** (Metadata Encoding and Transmission Schema): The METS schema, is a flexible XML framework designed for storing administrative, structural, and descriptive metadata about digital objects in a digital library. In addition to encapsulating the metadata itself, the framework provides elements for describing the relationship among the metadata and among the pieces of the complex objects http://www.loc.gov/standards/mets/
Data and metadata exchange	Metadata Data Information Packages	Syntactic/Semantic	**CCSDS-XFDU** The XML Formatted Data Unit (XFDU) standard is being developed by CCSDS. An XFDU is a logical unit made up of a primary Package Interchange File (PIF), also known as the XFDU package, plus any other files, PIFs or repositories referenced by the primary PIF. The PIF is a physical container file such as a ZIP or TAR file, containing an XML manifest document and none, some or all of the files referenced by the manifest http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206610R1/Attachments/661x0r1.pdf
Data and metadata exchange	Metadata Digital content	Syntactic/Semantic	**BagIt** is a hierarchical file packaging format for the exchange of generalized digital content. A "bag" implements a structure to enclose descriptive "tags" and a "payload" but does not require any knowledge of the payload's internal semantics. This BagIt format should be suitable for disk-based or network-based storage and transfer. https://wiki.ucop.edu/display/Curation/BagIt
metadata exchange between Current	Metadata for publication information	Semantic	**CRIS-OAR format** is a metadata exchange format for publication information with an associated common vocabulary.

Interoperability issue/challenge	**Interoperability objects**	**Interoperability level**	**Adopted solution/ model/ standard /service**
Repositories Information Systems and Open Access repositories			http://www.knowledge-exchange.info/Default.aspx?ID=340
Defining a standard for a harvestable set of descriptive metadata	Metadata about museum objects	Semantic	**LIDO (Lightweight Information Describing Objects):** is an XML harvesting schema supporting the full range of descriptive information about museum objects in a standard way. http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/
Sharing metadata between services metadata harvesting: combining metadata from different repositories into a combined data store	Metadata	Semantic	The **Open Archives Initiative Protocol for Metadata Harvesting** (OAI-PMH) is a lightweight protocol for harvesting records containing metadata from multiple repositories. The OAI-PMH gives a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). The metadata that is harvested may be in any format that is agreed by a community (or by any discrete set of data and service providers), although unqualified Dublin Core is specified to provide a basic level of interoperability. Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested, or “aggregated” data. The link between this metadata and the related content is not defined by the OAI protocol. It is important to realise that OAI-PMH does not provide a search across this data, it simply makes it possible to bring the data together in one place. In order to provide services, the harvesting approach must be combined with other mechanisms. http://www.openarchives.org/pmh/
Cross-system content transfer	Digital resources	Syntactic/ Semantic	**SWORD** (Simple Web-service Offering Repository Deposit) is an

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
			interoperability standard that allows digital repositories to accept the deposit of content from multiple sources in different formats (such as XML documents) via a standardized protocol. The project aims were: • To improve the efficiency and quality of the repository 'Ingest' function • To diversify and expedite the options for timely population of repositories with content • To facilitate the creation and use of common deposit interfaces • To improve repository interoperability as outlined in the Information Environment • To take a service-oriented approach to development as outlined by the E-Framework SWORD is a profile (specialism) of the Atom Publishing Protocol, but restricts itself solely to the scope of depositing resources into scholarly systems. http://www.ukoln.ac.uk/repositories/digirep/index/SWORD_Project
			The Open Access Repository Junction API is an API developed in the context of the OA-RJ project with the aim of supporting authors to deal with automatic deposit of multi-authored, multi-institutional research articles into multiple repositories. The OA-RJ API uses machine processing to refer and redirect users to the most appropriate repositories by determining depositors' institutional affiliation and associated repositories. An associated functionality, named ORI (Open and Repository Identification) deals with identification of academic organizations and associated repositories. http://edina.ac.uk/projects/oa-rj/index.html

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
Defining a common way to identify and describe aggregations of Web resources (e.g. overlay journal issue; a multi-page html document; a collection of images and so on)	Aggregations of Web resources	Syntactic/Semantic	**Open Archives Initiative Object Reuse and Exchange (OAI-ORE)** defines standards for the description and exchange of aggregations of Web resources. http://www.openarchives.org/ore/
Providing a searchable web database of technical information about file formats	Technical information about electronic records: representation information	Technical/Semantic	**PRONOM** is a web-based file format technical registry, which captures detailed technical information on file formats, the software tools required to access them, and the technical environments required to access them. The registry has been designed to support future interoperability with other registry systems. In particular the **PRONOM Persistent Unique Identifier** (PUID) is an extensible scheme of persistent unique identifier adopted for enabling the access, exchange and management of representation information (i.e. the format) described in the PRONOM registry. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
Linking data distributed across the Web	Web resources	Syntactic/Semantic	**Linked Data Initiative:** describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF. http://linkeddata.org/home
DP Area: AUTHENTICATION/AUTHORIZATION/RIGHTS			
Alignment of authorization/authentication protocols	Identity information	Semantic	1) **OpenID** is an open decentralized user's authentication standard allowing users to log on different systems with the same digital identity http://openid.net/ 2) **Security Assertion markup**

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
			Language (SAML) is a standard which defines an XML based framework for exchanging security information between online business partners. http://saml.xml.org/
To create a standards infrastructure for rights management and licensing	Intellectual Property rights and licensing (identifiers, metadata, iconography and messaging)	Semantic	**Linked Content Coalition**: is a cross-media project that will create the framework for a fully interoperable and fully connected standards-based communications infrastructure so that businesses and individuals can manage and communicate their rights more effectively online. http://www.linkedcontentcoalition.org/home_page.html#!home/mainPage
To create an infrastructure for rights management	Rights information	Semantic	**ARROW system:** developed during the ARROW project (Accessible Registries of Rights Information and Orphan Works) is a distributed infrastructure including the following components: the ARROW Web Portal Services, the Rights Information Infrastructure (RII), the ARROW Work Registry (AWR) and the Registry of Orphan Works (ROW). The system has been thought as an interoperability facilitator addressing several interoperability issues: - data interoperability at the transnational level - interoperability between catalogues of libraries, catalogues of books in prints and repertoires of the Reproduction Rights Organizations (RRO); - data created at the book level (manifestation) and rights defined in terms of works (expression). http://www.arrow-net.eu/presenting-arrow-system
DP Area: PRESERVATION MODELS and SERVICES			
To develop a conceptual	Archival concepts,	Technical Semantic	**OAIS Reference Model:** is a high level conceptual framework for an Open

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
framework for the standardization of archiving systems	terminology, functionalities, strategies and techniques	Organizational Political	Archival Information System (OAIS). It defines common terminology and concepts of archival concepts needed for long-term digital preservation and comprises several models (functional, information, and environment), which prescribe a minimal set of responsibilities required for the preservation of digital information. The framework focuses also on interoperability issues related to the interaction and cooperation among archives, distinguishing different levels of interaction (i.e. independent, cooperating, federated and shared resources archives. See Section 6 of the document available through the following link). http://public.ccsds.org/publications/archive/650x0m2.pdf
To make preservation standards and tools interoperable	Applications and services for preservation and data repositories	Technical Semantic Organizational	**Planets Interoperability Framework:** is a software infrastructure for the integration of new and existing preservation services. The framework is based on Java Enterprise Edition standard (Java EE 5) for the implementation of Web services and Web applications. The interoperability challenge is addressed by: - enforcing a set of standard Web service profiles for preservation services; - providing a common model for the digital objects on which preservation actions are carried out; - defining atomic preservation actions such as identify, characterize, compare, modify, migrate and view; - defining and implementing a preservation workflow and related services. http://www.planets-project.eu/docs/presentations/Interopera

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
			bility_Framework_RK.pdf
Common preservation framework for heterogeneous data	Digital preservation functionalities implemented by software components	Technical Syntactic Semantic Organizational	**CASPAR framework** is a software platform (based on an extension of the OAIS information model) that integrates advanced components to be used in a wide range of preservation activities and enables the building of services and applications that can be applied to multiple areas and adopted by different user communities. http://www.casparpreserves.eu/caspar-project.html
Shared Digital preservation framework and tools	Digital preservation functions and processes	Technical Syntactic Semantic	**The SHAMAN framework** is a next generation (based on the technology infrastructure of data grids) integrated framework for digital preservation including systems and tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives. http://shaman-ip.eu/node/10
Management and sharing of distributed information	Microservices and rules for interoperability	Technical Syntactic Semantic Organizational Legal	**iRODS (Integrated Rile-Oriented Data System)** is an open source software system for key data management tasks which supports management, sharing, publication and long-term preservation of distributed data. The specific interoperability challenges addressed by iRODS are: - Management of interactions with storage resources that use different access protocols. - Support for authentication and authorization across systems that use different management systems. - Support for uniform management policies across institutions that may have different access requirements. - Support for wide-area-network access. https://www.**irods**.org/

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
DP Area: RESEARCH DATA DEPOSIT, DISCOVERY, ACCESS, REUSE AND CITATION			
to make better use and reuse of research data outputs	Research data	Technical Semantic Organizational Inter-community	**Australian National Data Service:** is an Australian initiative which aims to create a cohesive national collection of research resources and a richer data environment that will: • to make better use of Australia's research outputs • to enable Australian researchers to easily publish, discover, access and use data • to enable new and more efficient research http://www.ands.org.au/
Durable access to research data (mainly in the Social Sciences and Humanties	Research data	Technical Semantic Organizational Inter-community	**EASY (Electronic Archiving System)** Trusted Digital Repository of DANS (Data Archiving & Networked Services) (http://easy.dans.knaw.nl) Depositing research data (self-depositing) Retrieval of research data Archival storage License management
to discover and reuse research data from repository		Technical Semantic Organizational Inter-community	**SRU Search/Retrieval via URL** SRU is a standard XML-focused search protocol for Internet search queries, utilizing CQL (Contextual Query Language), a standard syntax for representing queries. http://www.loc.gov/standards/sru/index.html
To create research networks for linking researchers and publications	Publications, annotations, tags, document usage data and other metadata	Semantic Organizational Inter-community	**Mendeley** is a desktop and web platform for managing and sharing research papers, discovering research data and collaborations. http://www.mendeley.com/
To support the identification, location and reuse of information	Digital collections	Semantic Organizational Inter-community	**Recollection** is a platform for enabling communities to support the distributed curation of digital content created in partnership with the U.S. Library of Congress and the National Digital

Interoperability issue/challenge	Interoperability objects	Interoperability level	Adopted solution/ model/ standard /service
across distributed digital collections.			Information Infrastructure and Preservation Program (NDIIPP). Recollection aims to facilitate data sharing and customized display of data in an arbitrary social environment. The platform leverage the Web architecture Semantic Web technologies and social network practices to facilitate content access, sharing and integration with other digital information sources. http://viewshare.org/
To facilitate search, access and reuse of bibliographic records in agricultural science and technology.	Bibliographical records	Semantic International Inter-community	**AGRIS** (International System for Agricultural Science and Technology) is a global public domain database, maintained by FAO, with more than 4 million structured bibliographical records on agricultural science and technology. Each record includes metadata, such as conferences, researchers, publishers, institutions and subjects, catalogued from more than 150 participating institutions in more than 100 countries. The AGRIS Search system, accessible at http://agris.fao.org, allows scientists, researchers and students to perform sophisticated searches using keywords from the AGROVOC thesaurus, specific journal titles or names of countries, institutions, and authors.
Author identification	Authors (profiles) and the records of their works	Semantic International Organizational Inter-community	**AuthorClaim registration service:** is an application, which contributes to the identification of authors (editors and potentially other contributors) and link them with the records about the works that they have written, as recorded in a bibliographic database. AuthorClaim provides a service where authors can register, leave personal data such as their names, affiliations and homepage URL. Then they can claim to be the author of document described in some contributing bibliographic databases. http://authorclaim.org/about

The analysis of the projects and initiatives in Section 2.4 combined with that of the current interoperability solutions presented in this section provide a synopsis of the current effort on interoperability in the landscape of digital preservation. In Figure 16 we present a graphical representation of this result. Six main areas of investigations have been identified: 1) Metadata; 2) Persistent Identifiers; 3) Data Infrastructures for e-science 4) Scientific Data services 5) Linked Data and Semantic Web technologies 6) Preservation services. For each area we have reported the main investigation topics that we have encountered in the analysis.

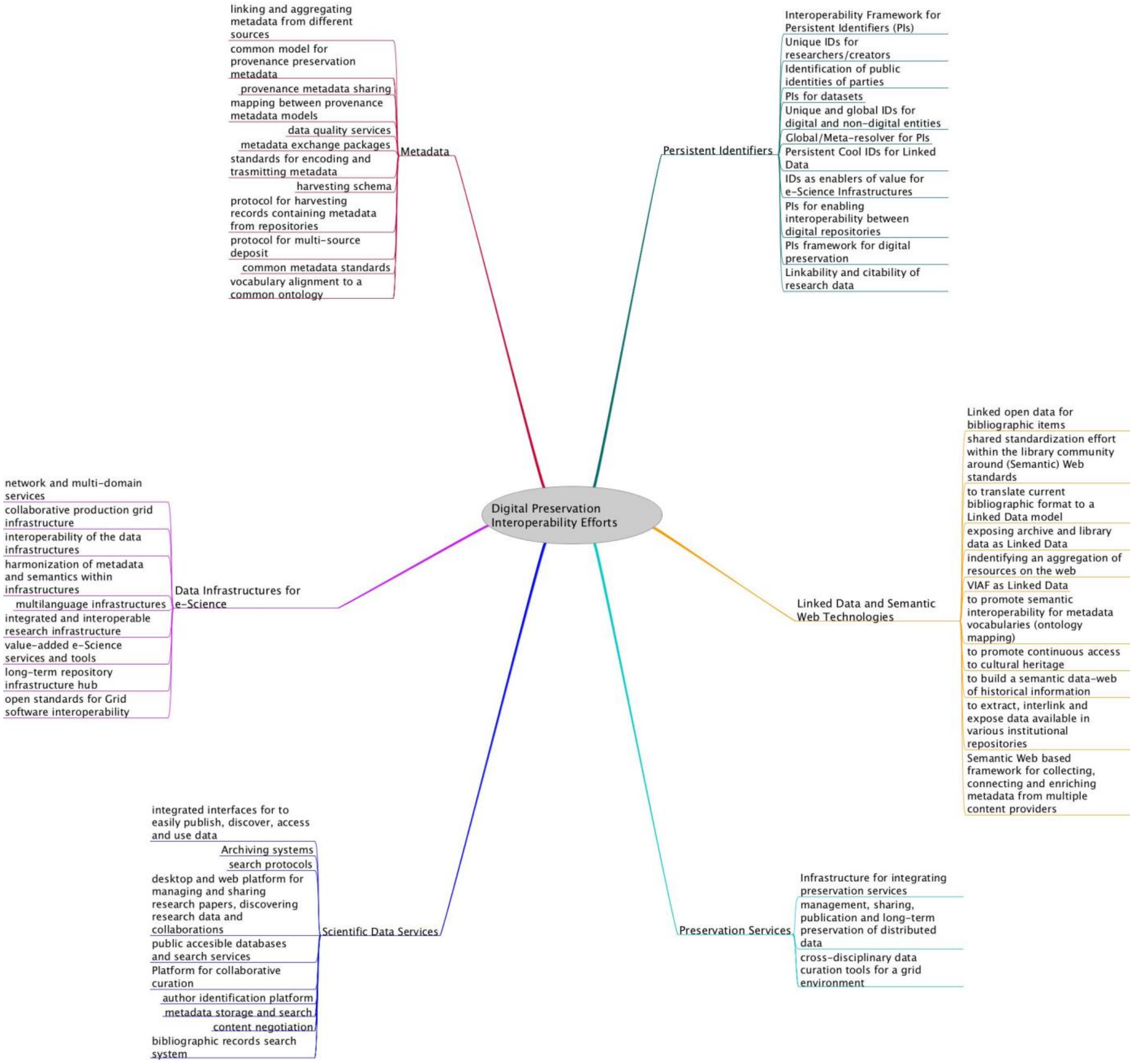


Figure 16: Interoperability efforts in the DP Landscape

5 INTEROPERABILITY OBJECTIVES AND GUIDELINES

In this section we translate the insight gained from the disciplinary scenarios (and related challenges) and the analysis of interoperability solutions and services, into common interoperability objectives and guidelines.

5.1 CURRENT SITUATION AND IDENTIFIED GAPS

As it emerges in the first part of this work, there are a number of gaps between the current situation and the desirable objectives, scenarios and services, which would be enabled if effective interoperability solutions were in place (as illustrated by the example scenarios described in Section 3). In this section the gaps are diagnosed identifying the current situation (where we are), the future goals and objectives (where we want to be) and the possible solutions to fill the gaps (how to fill the gap). An example of the results of the analysis in the domain of PIDs is graphically reported in

Figure 17.

On the one hand, the output of the analysis can be used to raise awareness about the weaknesses of the current situation and give directions to future actions that need to be taken. On the other hand, the output of the analysis can be useful to identify possible opportunities, which can be exploited to make these actions more effective.

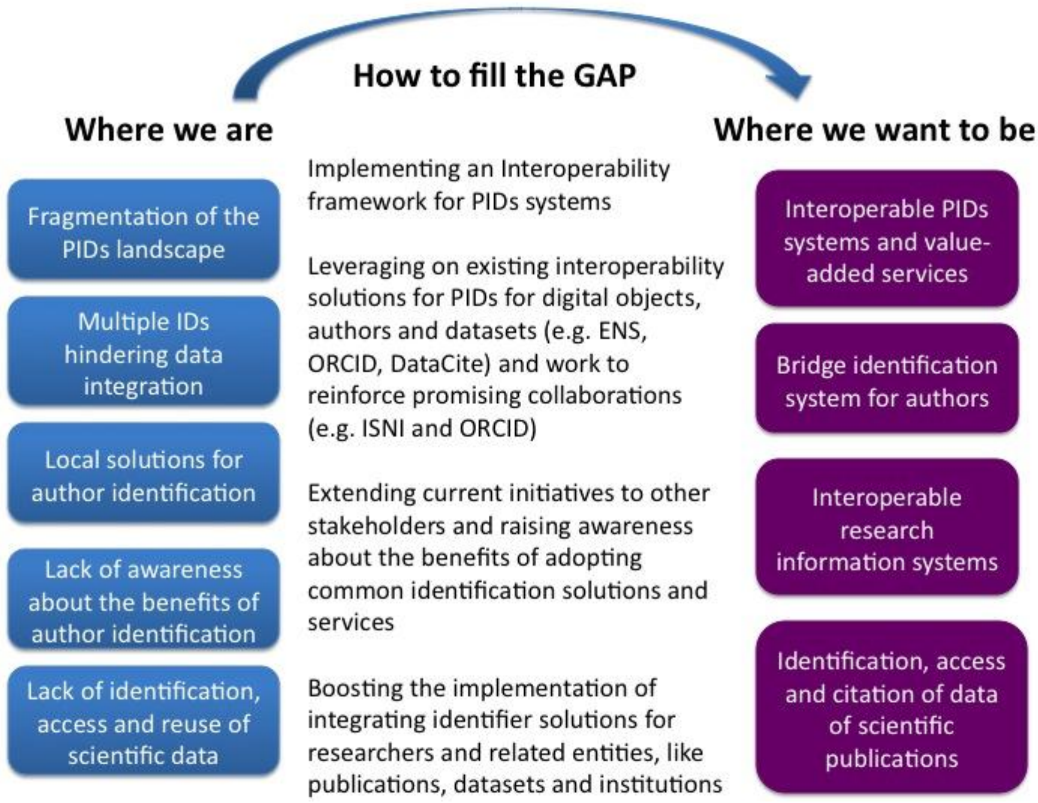


Figure 17: An extract of the Gap Analysis results in the PIDs domain

Where we are	Where we want to be	How to fill the gap
Digital Objects Identifier Systems		
Fragmentation of Persistent Identifiers (PIDs) domains in identification approaches, policies, business models, objectives, semantics, etc. Registration Agency are not aware that different PIDs are associated to the same object representation during its lifecycle. Therefore, the same digital object is usually identified by multiple identifiers in different systems and in different moments of its lifecycle.	Enabling the reliable long-term access to integrated information using Persistent Identifiers through a single entry point which make different PID solutions interoperable.	- Implementing the APARSEN Persistent Identifier Interoperability Framework. - Leveraging on existing solutions for managing unique identifiers for digital and non-digital entities, like the OKKAM ENS or EZID, by developing value added preservation services on top of their technological infrastructure.
Author/Creator Identifier Systems		
The authors / creators identification is currently managed through authority file sometimes implemented at organization level or international level as VIAF. The Persistent Identifier systems for authors are scarcely adopted and there is a fragmentation of the landscape of PIDs systems that are also used within territory boundaries (local ad-hoc author identification systems).	A bridge identifier solution connecting proprietary and local author identification systems. Exploiting multiple persistent identifiers association to a single person assigned across boundaries (roles, domains, etc.) to enhance information integration from different sources and provenance assessment. Developing interoperable research information systems capturing and exploiting much richer relationships between the author and its related entities (institutions, collaborators, publications…) to assess the quality and the impact of his/her work.	- Boosting the implementation of integrating identifier solutions for Authors/Persons such as ISNI and/or ORCID and their integration. - Extending current initiatives to other stakeholders and raising awareness about the benefits of adopting common identification solutions and services. - Extending the Entity Name System platform developed by OKKAM as an interoperability solution for Persistent Identifiers for authors and other kinds of entities, like data and institutions. - Enriching library authority files by Linked Open data sources through the representation of authority files as LOD and link them to LOD related resources.

Where we are	Where we want to be	How to fill the gap
Datasets Identifier Systems		
Currently, the majority of datasets, independently by their domain of production, are difficult to be accessed, cited and shared. This is due to some peculiarities of datasets like granularity (which elements within the datasets are being referred to), versioning (in case of dynamic datasets, which version is identified). There is a lack of clarity and agreement on what "authorship" of a dataset means. There is no agreement on what persistence and longevity of data is required for it to be citeable and cited.	Uniquely identifying, describing and accessing data that is referenced in scientific publications. Identifying and acknowledging the creators of datasets.	- Extending the DataCite initiative, for example by reinforcing the collaboration with associations of publishers. - Supporting the Linked Open Data initiative to enable open-access data. - Strengthening the links between data citation and author identification systems. - Raising the awareness in the community about the benefits of data citation for digital preservation.
Library Classification Systems		
Several library classification systems are currently in use such as Dewey Decimal Classification, Library of Congress Subject Headings. Others are built within national and/or domain boundaries.	Identifying entities across systems, which can use different vocabularies to classify them.	- Automatic mapping between different categorization systems and thesauri to implement search services. - Publishing controlled vocabularies as Linked Data and exploiting Semantic Web technologies to create mapping between them and linking the represented contents.
Library Linked Data (LLD)		
Library linked data is a reality and several notable efforts have been done, like for example, Europeana, id.loc.gov, VIAF as linked data, DNB, Library Framework Initiative. However several limitations and challenges prevent LLD full potential:	Interlinking different digital objects and representations coping with heterogeneity of formats, language and vocabularies. Multilingual services and cross-lingual linking and data integration.	- Defining methods, techniques and tool to deal with heterogeneous source formats, schemes and encodings. - Involve library experts in developing mapping across metadata schemas. - Promoting the development

Where we are	Where we want to be	How to fill the gap
- Lack of methods to deal with heterogeneous formats. - Heterogeneous metadata standards in use. - Lack of mapping across vocabularies. - Lack of multilingual vocabulary elements. - Lack of cross-lingual mechanisms to link resources across libraries. - No extensive use of mechanisms to indicate provenance, license, rights. - Lack of end-user services.	End-user interfaces providing enriched information spaces. Integrating LLD into the library workflow (curation, cataloguing).	and use of multilingual cataloguing dictionaries (e.g. multilingual IFLA namespace). - Promote cross-library collaborations for linking resources across system and organization boundaries. - Improving linking at bibliographic level. - Developing scalable infrastructures to facilitate consumption of Linked Data. - Developing documented APIs to provide access and search over Linked Data.
Metadata		
Content providers manage their resources with a number of different metadata sets, different complexity, refreshing rate, stability, etc., according to the domain requirements. A plethora of metadata standards and formats have been developed to support the representation and management of digital objects. The diversity of metadata standards, the existence of local schemas and the heterogeneity in metadata usage and implementation has significant implications for institutions to provide integrated access to information resources when they attend to share contents across their system and organization boundaries. Moreover, the existence of several metadata standards, coupled with the proliferation of in-house schemas make the adoption of a unique standard, as a possible solution to allow interoperability, a daunting task.	Making semantic connections between Web entities and related content in order to facilitate searching and finding structured results also across multiple systems, domains and languages.	- Metadata integration via linked data solutions (e.g. Metadata Management System that uses Semantic Web technologies to handle the problem of heterogeneity of metadata) - Aggregating diverse data sources and performing vocabulary alignment to a common ontology
Metadata registries are "formal systems that can disclose	The implementation of metadata registries as part of a digital	RDF and Semantic Web technologies, which have been

Where we are	Where we want to be	How to fill the gap
authoritative information about the semantics and structure of the data elements that are included within a particular metadata scheme" (Heery, Gardner, Day, & Patel, 2000). They typically store the semantics of metadata elements in a specific domain (e.g. the Environmental Data Registry (EDR) provided by the US Environmental Protection Agency) or for a specific type of content (e.g. The SMPTE (Society of Motion Picture and Television Engineers) Metadata Registry for audiovisual content), and provide mappings to other metadata schemas.	preservation system may enable the management and re-use of metadata and facilitate the exchange of metadata and information between repositories and other preservation systems. In particular a metadata registry as a component of a preservation system should support the following basic functionalities: • Providing authoritative information about the digital preservation metadata terms and vocabularies used in the system; • Mapping the metadata schema adopted by the system to other metadata schemes or versions of the same schema. Both functionalities could help the implementation of automatic services for data ingestion and export, like facilities for metadata conversion.	already adopted to develop metadata registries, may also provide a valuable solution for developing metadata registries in digital preservation context.
Ontologies and Vocabularies		
Several ontologies and thesauri are currently in place in the Earth Science domain, such as INSPIRE GEMET Ontology and GSDCA EO Ontology.	Discovering data systems via controlled vocabulary, which would permit the user to discover resources through familiar terminology in the EO domain.	- promoting and adopting as much as possible the OGC standardization initiative - Providing an Ontology mediator reuse and merge ontologies either complementary to each other, or partially/totally redundant
Data Provenance		
There are several models for modelling and recording provenance, and various mappings between these models but the institutions are slow in adopting them.	Common vocabularies and data models are needed to express basic provenance information in an interoperable fashion. Exchanging and aggregating provenance information of various processing tasks or transfer/archiving events.	- Supporting the adoption of provenance models and exploiting the mappings that have been established between such models (also the mappings defined in the context of APARSEN/WP24) for the digital entities. This however presupposes the ability to give persistent identifiers to the involved entities (actors,

Where we are	Where we want to be	How to fill the gap
		processes, artifacts, etc).
Preservation Tools		
Many different suites and preservation tools are in use in different communities (e.g. iRODS, LOCKSS). The isolation from each other represents an obstacle for inter-institutional interoperability.	To develop a more comprehensive suite of interoperable tools and allow the digital preservation community to leverage the power of each system in modular fashion.	- To develop tools and methods needed to automate the exchange of large collections of data between different preservation systems.
Exchange Standards		
There is a wide agreement in the international preservation community that responsibility for long-term preservation of scientific and cultural heritage materials must be shared among many organizations. A direct implication of this is that there are many use cases for transferring copies of stored information packages from one repository to another. Currently different repositories use different applications, which use different storage packaging, require different processing information, supply different metadata and may implement different preservation strategies.	To implement solutions (like an international standard for Package Transfer), for enabling interoperability between preservation repositories, i.e. enabling content from one preservation repository to be ingested by another preservation repository.	- to promote the development and use of a common exchange package based on existing standards widely adopted by the preservation community.
Preservation Framework		
Most digital preservation strategies depend upon the capture, creation, maintenance and sharing of preservation metadata including information about provenance, authenticity, preservation activity, technical environment, rights management which make objects self-documenting across time. A number of organizations have developed metadata of this type in support of their own preservation activities. The lack of cross-organization coordination of this effort led to the development of a set of metadata reflecting the particular needs and	Digital libraries, repositories, publishing systems and other systems need a common, continuously updated metadata framework for digital preservation that met the general consensus of the involved communities, and could be readily applied to a broad and growing range of digital preservation activities. The framework should support not only the metadata creation but also the metadata exchange between repositories and other preservation systems.	- Improving the implementation and use of the PREMIS Data Dictionary for example by: ○ gathering PREMIS-related vocabularies into central registry to promote convergence on common standard vocabularies. ○ working on the integration of the PREMIS dictionary and schema with Metadata Encoding and Transmission Standard (METS).

Where we are	Where we want to be	How to fill the gap
requirements of the specific community that authored them. A first effort toward the development of a common integrated metadata set has been done by the OCLC/RLG working group on preservation metadata that focused on the collaborative development of a preservation metadata framework. Based on these initial prototype metadata elements, the PREMIS working group developed a data dictionary of core metadata for archived digital objects, as well a set of best practices for creating, managing, and using the metadata in preservation systems.		○ identifying areas of the Dictionary that need to be improved or expanded due to the emergence of different needs and requirements for the preservation of new types of contents. ○ enabling the exchange of PREMIS-conformant metadata between repositories by addressing the issues involved in extracting and exporting PREMIS metadata from one digital archiving system in ways that ensure it can be received, ingested, and understood by another archiving system. - developing automated tool for creating, managing and processing preservation metadata.
A key problem that digital preservation is currently facing is the management of growing volumes of digital data, which are distributed across national, organizational and system boundaries. Currently, scalable large-scale and cross-boundary preservation infrastructures are not widely diffuse and local solutions are still the predominant approach.	1) a scalable infrastructure for the efficient planning and application of preservation strategies for large and heterogeneous data collections and 2) a collaborative preservation framework which allow local preservation systems to interoperate across organizational, national and systems boundaries.	Both requirements can be addressed by implementing a large-scale collaborative preservation framework, based on collaborative storage and coordinated data sharing and long-term management. The CASPAR high-level model of virtualization, discussed in Section 2.1 could represent the starting point for the development of a large-scale platform enabling the building of services and applications that allow preservation systems to plan collaborative preservation actions.
Semantic Annotation Services		
Current legacy cataloguing	To create Cultural Heritage	Domain independent Semantic

Where we are	Where we want to be	How to fill the gap
systems in memory organizations do not support creation of ontology-based annotations. If semantic annotations cannot be created in memory organizations already when cataloguing content, then costly manual work is needed later on when transforming and disambiguating literal metadata for shared portal use.	information portals where heterogeneous contents are semantically interoperable and mutually interlinked.	Web standards can be used to create a public domain infrastructure of ontologies which can be made operational by ontology and annotation services that can be connected easily with cataloguing systems.
E-science Infrastructures		
E-science infrastructures are emerging in Europe and world-wide. However, these efforts are today highly fragmented due to a lack of global coordination mechanism, resulting in domain-, discipline-, institution-, and country/region-specific implementations that are not interoperable. This lack of interoperability decreases the value of investments in data infrastructure since each investment may not benefit from others. It also increases the costs of data preservation, discovery, access, and re-use and re-purposing by preventing automated solutions and limiting economies of scale.	To define an international framework for interoperable Data Infrastructures for e-Science which includes a preservation layer for enabling data re-use and sharing.	- To involve all stakeholders in the scientific process in the design of the framework. The definition of the interoperability framework for e-science infrastructures could benefit from global cooperation and coordination activities like the DataWeb forum (described in Section 2.4.2). - Exploiting the results of projects like ODE project, to identify 1) issues, needs and requirements of relevant stakeholders, 2) best practices and solutions in data sharing, data re-use and citing and 3) the drivers for the change and barriers impeding progress.
Research Information Services on a national scale, like NARCIS[22] in the Netherlands or LATTES[23] in Brazil, have started to become much popular in recent years providing a unique gateway to national scholarly information (publications, data sets, thesis, researchers, organizations). A central aspect of the interoperability within the national research management	Collaboration and sharing data between research information systems to avoid duplication of effort is necessary. Therefore a single point of entry to the research data managed by these systems should enhance the accessibility and visibility of the research data contributing to the worldwide scientific advancement.	Interoperability across national research management systems can be enabled by adopting a unique global identification solution (or by ensuring interoperability between local identifier systems) to ensure the persistent accessibility to distributed content. The OKKAM ENS or PIDs interoperability solutions, like the APARSEN interoperability framework for PIDs systems or ORCID, could become an

[22] http://www.narcis.nl/
[23] http://lattes.cnpq.br/

Where we are	Where we want to be	How to fill the gap
system is the use of unique identifiers to interlink content. For example in NARCIS, the data from institutional OAI-PMH repositories (Irs) and the information from the Current Research Information Systems (CRISs) in the Netherlands are interlinked by identifiers such as the Digital Author Identifier (DAI), a unique identifier assigned to each researcher in the Netherlands.		essential component for the development of a cross-national research information service.

Table 4: Gap Analysis Results

5.2 RECOMMENDATIONS

In order to put theory into practice we have devised two sets of recommendations, which should promote the realisation of interoperability in the areas of long-term preservation.

The first set includes general recommendations that are applicable to all the categories of stakeholders and aim at:

- Fostering the broad adoption of common standards and specifications reducing dependencies, facilitating the interoperation between systems for the entire digital object lifecycle management process and enabling higher-level services on top of standard compliant systems.
- Promoting the use of appropriate identification systems and their interoperability.
- Promote the convergence toward common policies and governance models, which favour the adoption of interoperability solutions and trust on them.
- Ensuring the necessary long-term financial support and the efficient use of economic resources.
- Raising awareness and securing trust among the involved stakeholder communities.

5.2.1 General recommendations

5.2.1.1 STANDARDS

- **Standards are Good.** One recommendation is to rely on **Standards** in case there are appropriate standards for the digital objects at hand (apart from the standards mentioned in the previous sections, a long list of standards is given in WP13, D13.1). If compliance to one standard guarantees the achievement of one or more interoperability objectives, then the adoption of the standard is certainly beneficial. From a dependency point of view, we can say that the standardization essentially makes the *dependencies* more clear and resolvable. For instance consider a digital object DO. Instead of relying on X proprietary components for rendering it, it is better to rely on the tools that are available for a particular standard Y.

- **Standards are not a Panacea.** We should be however aware that standardization does not vanish the dependencies of the digital objects, e.g. if a standard Y becomes obsolete and there are no longer tools that support it, then the object DO is in trouble. A rising question is whether we could tackle the interoperability problem without having to necessarily rely on several and possibly discrepant standards. Can we come up with processes that can solve (or aid) the interoperability problem in a flexible manner? What kind of model/services could support this? Can we exploit ***emulators*** and ***converters*** to reduce our dependencies or to tackle the problems of vanishing or evolving standards? One way to approach these questions will be investigated in the context of Task 2520 (Intelligibility Modelling and Reasoning) and the results will be reported in the second deliverable of WP25: D25.2 (due August 2013). Some preliminary results are described in (Tzitzikas, Marketakis, & Kargakis, 2012).

- **Define Interoperability Standards through the entire lifecycle of a digital object.** Standards should regulate the entire chain of digital preservation steps that form the lifecycle of a digital object from its creation to its re-use through the process of digital preservation.

- Content holders should support and deploy **architectures, protocols and standards** to ensure that their digital content interoperates with other services and collections. Architectures must be sufficiently flexible to accommodate digital content from a variety of sources (including different types of cultural heritage organizations, individuals and communities), and should provide seamless access for end users while preserving provenance information for documentation purposes.

- **Involve stakeholders in the definitions of standards**. Since it is difficult to mandate standards, it is easier to work on community accepted standards. Community evolution of standards should be encouraged. A concrete example of a successful coordinated effort between two communities (library and museum) to define a common interoperability standard (FRBRoo ontology) to contribute to the solution of the problem of semantic interoperability between their documentation structures is shown in Figure 15.

5.2.1.2 IDENTIFICATION

- **Bootstrap an interoperability solution for Persistent Identifiers**. The persistent identification of digital objects (e.g. articles, datasets, images, stream of data) and non-digital objects (namely real-world entities, like authors, institutions but also teams, geographic locations and so on) is becoming a crucial issue for the whole information society and for the development of e-Science infrastructure in particular. However the proliferation of several PIDs systems within different communities and the resulting fragmentation of the PIDs ecosystem pose the key challenge of establishing an interoperability solution among the current PI systems to enable the persistent access, reuse and exchange of information across different systems, locations and services.

  Therefore, actions are needed to bootstrap the convergence toward an interoperability solution for PIDs which open new prospects for advanced value added information integration services. Hover, since any identifier system is always used within cultural, organizational, geographical and disciplinary boundaries through a technical system, it follows that devising an appropriate solution to the problem of identifiers interoperability is far from being a merely technical issue. This means that any action to bootstrap an interoperability solution needs to work towards systematic implementation of those organizational, political, social and economical factors that foster trust and agreement among the relevant stakeholders.

- **Elaborate on Information Identity.** Apart from the problem of identifiers, another critical point is the identity of the content. Even though library and archival practice, as well as Digital Preservation, have a long tradition in identifying information objects, the question of their precise identity under change of carrier or migration is still a riddle to science. One theory, developed in the context of APARSEN, that tries to give some light to this aspect is described at (Doerr & Tzitzikas, 2012). The objective is to provide criteria for the unique identification of some important kinds of information objects, independent from the kind of carrier or specific encoding. The approach is based on the idea that the substance of some kinds of information objects can completely be described in terms of discrete arrangements of finite numbers of known kinds of symbols, such as those implied by style guides for scientific journal submissions.
  The basic thesis is that a notion of identity of information object, that conforms with legal practice and the intuition of Digital Preservation, must be based on an analysis of the intended sensory impression rather than the binary form or material embodiment of an information object. If the information object we are interested in is defined in terms of a finite, discrete arrangement of symbols, where each symbol belongs to a finite symbol set and adequate arrangement rules, it is then possible to extract from the sensory impression a symbol structure which is the substance of the carried information object. For that purpose (Doerr & Tzitzikas, 2012) proposed an ontology, actually a structurally object-oriented conceptual model that refines ISO 21127:2006 (CIDOC CRM), that provides the concepts and definitions necessary to demonstrate the feasibility of objectifying the content of a sensory impression to a particular *intended* symbol structure.
  An interesting consequence of this model is that an information carrier in general carries more than one information object - depending on the definition of the relevant features applied - and that some of those potential information objects may *incorporate* several other meaningful information objects with less features. This is in contrast to current Digital Preservation Research, which assumes one information content for a material or digital document. Information has a function it serves communication, and hence has to do with intentions. Therefore it appears quite natural that the proposed model suggests that the definition of information content depends on intention. Objective definition of information content in a use-neutral way appears to be impossible. The authors of that work argue that some special applications in literature studies, which may analyze any feature on an information carrier, such as spelling errors or stress of manual writing strokes, are of forensic nature, and should not be confused with information as an object of social function, but are frequently used as argument, that information content cannot be identified. These considerations may explain, why Digital Preservation Research has so much difficulties to identify what has to be preserved.
  As examples of practical applications, creators of information would like to define which are the contents that when preserved ensure the identity or authenticity of their work. Curators and archivists would like to record formally the decisions of what has to be preserved over time and to decide (or verify) whether a transformation preserves the intended features. In order to use the introduced model in practice, one would be need to define the information formats and respective symbol structures for a series of relevant functional domains, such as scientific publishing in computer science, etc. An extensible system of *feature ontologies* is expected to emerge from such an activity, which could be used by authors and Digital Preservation teams to select for specifying intended information content or content to be preserved.

5.2.1.3 ORGANIZATION, GOVERNANCE and TRUST

- **Raise agreement, increase awareness and social support.** Given the complexity of the interoperability exercise in many areas of digital preservation and the variety of stakeholders involved, a common direction must be defined. The involved parties should work together to define a common agenda ensuring a coordinated and interoperable digital preservation ecosystem. The VCoE of the APARSEN project should play a key role to coordinate the definition of this agenda due to its role in the creation of a common view and understanding about the preservation and interoperability requirements in different preservation domains and research areas. The definition of the agenda should also benefit of the results of other initiatives like the Opportunity for Data Exchange (ODE), which aims to capture opinions, experiences, best practices of a critical mass of relevant stakeholders about sharing, re-using, preserving and citing data.
  The agenda will define a clear conceptual framework, which will be a pre-requisite for dialogue and achieving consensus across the communities impacted, and serving as the basis for promoting awareness and mobilisation of skills and resources. The common agenda should include at least the following points:
  - Raising awareness about digital preservation interoperability objectives, challenges and available solutions.
  - Promote a cross-boundary view on challenging issues and opportunities.
  - Planning interventions to promote awareness, dissemination and education programs in order to reinforce knowledge and skills on interoperability strategies and solutions.

  The present deliverable can be used as a valuable instrument to address these points. First of all, the identified scenarios can be exploited to identify common interoperability needs and requirements and delineate priorities among them. Secondly the analysis of the current standards, models and services can be adopted to promote a general understanding of the available solutions and identify possible gaps and opportunities for future development. Finally, the contents of the document can be leveraged for training and dissemination purposes, by making some results visible. A brochure featuring the key messages of the deliverable or a dedicated Webpage continuously updated about the current initiatives and most diffused technological solutions may be envisaged to this purpose.

- **Foster good practice.** Spreading good practice for interoperability digital preservation needs to include a more deliberate exchange of lessons learned and case studies documenting the use of emerging solutions, workflows, and techniques across national, organizational and disciplinary boundaries. The analysis and evaluation of scenarios as well as the identification of prioritized interoperability challenges described in the present document can be used to benchmark available approaches and systems and identify best practices according to certain identified interoperability objectives. Moreover the use of specific variables of performance (e.g. sustainability of the solution, scalability) can be adopted to develop plans on how to make improvements and adapt specific best practices to specific contexts.

- **Promote and encourage coordination and collaboration among stakeholder communities around policies and governance.** The different needs and goals of the stakeholders involved in different areas of digital preservation may hinder the adoption of available interoperability solutions. Therefore, actions are needed to favour the convergence toward common policies and governance, which can help to achieve consensus across the communities. As described in Section 2.4 (see for example the Data citation Standards and Practice Task Group, the Library Linked Data Incubator Group or the International Group on FRBR/CIDOC CRM Harmonization), a possible way to encourage coordination and identifying collaboration tracks for the future may be to establish a board or task group of experts together with representatives from several communities with the aim of examining key issues, help coordinate activities and promote common models, policies and practices.

- **Work toward global trust.** Actions are needed to promote international agreement on global standards and policies. In this way, users can have evidence of authenticity for world-wide data (e.g. scientific) and resources. The creation of an European Framework for Audit and Certification of Digital Repositories[24] is an example of the actions promoted within APARSEN[25] to build global trust by enabling interoperability between increasingly challenging audit processes in digital preservation.

5.2.1.4 ECONOMIC

- **Devise sustainable solutions.** Securing long-term sustainability of an interoperability solution or service is a key factor for promoting its trust, adoption and success. This can be ensured only if the organization behind it is sustainable and can guarantee the longevity of the solution. This is not simply a matter of finding sufficient funds but concerns many different aspects.

- **Build a robust community behind the interoperability solution or service**. The first step to establish a sustainable interoperability solution is to gain the support of (possibly) all the involved actors. Interoperability solutions are only possible if cultural heritage institutions, governments, public administrations, research institutions and private organizations work in close cooperation in supporting them, sharing responsibilities and finding adequate business strategies. This is for example the case of ORCID (see Section 2.4.5) that can count on a broad community including many different stakeholders (like individual researchers, universities, national libraries, commercial research organizations, research funders, publishers, national science agencies, data repositories and international professional societies) and, prior to its launch, obtained the support of over 300 organizations, 50 of which provided start-up financial support to the initiative.

- **Align the interests, roles and responsibilities of the involved stakeholder communities into a sustainable economic strategy and operationalise them in a business model**. Again the stakeholder participation in the definition of sustainable business strategies is crucial. To this purpose a business working group including the representatives of all the communities can be created to review membership policies, budget models and investigating funding options to ensure the long term sustainability of the system.

- **Provide clear incentives to adopt the interoperability solution.** The lack of clear incentives to adopt a given interoperability service may threat the sustainability of the service. For example the adoption of shared methods and services by independent organizations may bring costs. Sometimes the costs are financial due to the purchase of hardware or software or for hiring and training staff. In other cases costs are organizational. Introducing a new standard requires inter-related changes to existing systems, altered workflow, changed relationships with suppliers and so on.

- **Maximize the return of investment**. Support research and development across organizational and national boundaries to identify interoperability solutions worldwide that yield the best

---

[24] http://www.trusteddigitalrepository.eu/Site/Welcome.html

[25] Other actions are discussed in the APARSEN brochure on Trust available at http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=APARSEN+brochure+on+Trust

return of investment.

5.2.2 Categorized recommendations

5.2.2.1 Recommendations to support interoperability for Persistent Identifiers (from the lesson learned in the DIGOIDUNA study and APARSEN WP22)

- **Define a common set of objectives and requirements among key stakeholders toward the design and implementation of an interoperability infrastructure for Persistent Identifiers**, taking into account technological, economical, social and political factors, which characterize the landscape of PIDs systems. The work done with the High Level Expert Group on Persistent Identifiers for the validation of the Interoperability Framework within the WP22 represents a first step in this direction, and has identified several requirements at different levels such as trust, policies, governance, persistence, sustainability and scalability[26].
- **Define and share a high-level framework** which describes the main entities involved, their relationships and the fundamental services as well, which can be used as a starting point for the design and implementation of a technological implementation of the interoperability framework. The Interoperability Framework for Persistent Identifiers systems, developed within the WP22 (see D22.1 for more details), is an example of a high-level conceptual representation which standardize the relationships between identified entities (e.g. digital objects, authors and institutions) and their Persistent Identifiers, creating a common layer where meaningful information about these entities stored by independent systems can be exchanged and integrated.
- **Build a community behind the interoperability infrastructure** to agree on a shared governance model, which devolves responsibilities among the involved parties, and define common trust criteria. Due to the fragmentation of the current landscape of persistent identifiers solutions and the different needs and interests of the stakeholders involved in different initiatives the actual implementation of an integrated solution can be hindered. Therefore, actions are needed to facilitate the convergence toward common policies on the governance of identifiers and toward integrating technical solutions.
- **Promote awareness and skills development** to enable different stakeholders to participate effectively on PIDs initiatives and infrastructures. The success of recent initiatives and events about Persistent Identifiers and their interoperability[27] demonstrates the intent of the main players and stakeholders to inform each other about recent development, user experiences and discuss trends and policies. Exchange opportunities of this kind, as well as more specific training and education initiatives, should contribute to mobilize the vast skills pool needed to implement interoperability infrastructures and services for Persistent Identifiers, by enlarging the user landscape to new key actors and creating a wider demand of the proposed solutions. To this purpose specific actions should be taken to spread the benefits of a systematic usage of identifiers interoperability solutions and value-added services on top of interoperable data and content.
- **Promote cross-fertilization between public and private sectors** to co-operate in the implementation of added value services on top of interoperability solutions. Funding bodies, for example, should support the development of collaborative models to create synergies

[26] A detailed analysis of the user requirements will be described in the D22.2.

[27] Just to mention some of them: Seminar on Persistent Object Identifiers, the Hague, 14 and 15- 06-2011; "Links that lasts", Cambridge, 19-07-2012; "International Workshop on Interoperability of Persistent Identifiers Systems", Florence 13-12-2012.

between scientific research production and the commercial sector. The public-private cooperation could stimulate more private investment, increase the knowledge sharing between academic and industrial research and strengthen the base knowledge of industry to develop service of public utility.

- **Foster interoperability based on consolidation of trusted and established PIDs systems**, like DOI, URN, ARK, instead of promoting the proliferation of ad hoc local systems.
- **Build sustainable business models** to guarantee the long-term sustainability of interoperability solutions. This can be obtained, for example, stimulating diversified investments (for instance from the private sector) complemented by targeted public investments to improve resource allocation and long-term sustainability. The flexibility of funding resources should enable the rapid scaling of promising solutions that embed or promote the value of existing identifier solutions.
- **Consider new types of interaction with and between structured data offered by the emerging Linked Data approach** and investigate whether PIDs and Linked Data can be integrated to create a new class of persistent interoperable cool URIs.

5.2.2.2 Recommendations to support semantic interoperability in the domain of earth sciences

It has been proved that, a widely agreed architecture concept, permitting the end-user to friendly discovery resources of interest, EO resources in this specific situation, is based on the use of ontologies, eventually mediated one another, easily and systematically linked to low-level metadata stored in specific metadata. Evidence of the effectiveness of this approach has been successfully provided within several ESA / EC founded and/or managed projects. Despite these promising demonstrators, most of the projects focused their attention on very narrow areas, providing very effective small-scale demonstrators. In the majority of the situations, these demonstrators partially or totally miss the capability of satisfying requirements at European or, even worst, worldwide scale.

To up-scale at European level such architecture concept, which currently only suits with local requirements, it is mandatory to involve the most relevant communities for both the two main actors of this scenario, end-users from one side and the EO resource providers on the other one. A fundamental prerequisite is that both the identified communities well represent the real scenario at European scale.

The end-user community is typically characterised by a deep expertise in specific application domains (e.g.: volcanology, forestry, etc…), but not necessarily in EO domain (e.g.: EO data distribution, sensors mode acquisition, etc…), on the contrary, the background of the resource provider community is mainly based on EO resources and pertinent technologies for providing them. The two parties need to be involved as much as possible in a synergic way, having them complementary roles during the definition, implementation and use of the architecture in object. From a very high level point of view the architecture is basically composed by three main components: ontology, ontology mediators and catalogue, each of them composed again by several sub-components. On the basis of the specific competence of the two communities, it is clear that:

- **the end-user community will have a key role within the definition of the ontology and mediators,** including:
  - Definition of dictionaries, starting from the state-of-the-art and real needs of such heterogeneous community
  - Definition of thesauri, which provide the semantically linked terms, collected within appropriate dictionary
  - Definition of ontology mediator, as needed, to semantically link available thesauri
  - Support in the standardisation activity concerning semantic representation language (e.g.:

OWL, SKOS, etc…)

- **the EO resource provider community will lead the definition and implementation of the catalogue technology facilities, including standardisation activities** needed to:
    - Seamlessly federate EO resources, sharing common Identity Management Services, Discovery Services, Invoke Services and Online Data (see HMA - http://wiki.services.eoportal.org/tiki-index.php?page=HMA%20Wiki)
    - Semantically annotate EO resources metadata, permitting to easily link ontologies to resource catalogue metadata

6 FUTURE WORK AND CONCLUSIONS

This deliverable discussed and analyzed the notion of interoperability and how it relates to digital preservation. It also analyzed the key questions about global semantic interoperability in digital preservation enabled by the Semantic Web initiative and Linked Data, including an overview of the main strengths and weaknesses of the approach, with a special focus on Cool URIs as an alternative solution to persistent identifiers for digital objects, authors and other entities relevant for digital preservation. In the sequel, it gave an overview of the current projects and initiatives on interoperability in different areas of digital preservation, by providing information about more than 60 projects and initiatives. Thanks to the experience of ESA, special focus has been given to projects and initiatives in the Earth Science domain that allowed to investigate semantic interoperability issues and challenges which are common to many other domains, such as ontology mapping, vocabulary alignment, multi-domain thesauri and vocabularies, metadata sharing. Then it described the main challenges encountered by partners and other stakeholders in their daily life activity by providing 13 interoperability scenarios evaluated according to three dimensions to provide an indication of the level of relevance of the challenges.

Based on the above, a matrix of models, standards and services for interoperability that cross the main areas of digital preservation which can be used as a tool to navigate the complex ecosystem of the current interoperability solutions was formulated. The matrix is clustered by DP areas and it contains 58 rows.

Based on the objectives and challenges addressed by the current solutions, and using feedback from the partners, the deliverable identified the main gaps between the current situation and future interoperability requirements that must be met and proposed possible strategies to fill these gaps.

Finally, the deliverable provided a list of recommendations and guidelines for ensuring interoperable digital preservation services, which can be used by the VCoE to define the Digital Preservation agenda for the future. These results will feed other work packages and tasks of APARSEN, mainly WP11 (Common Vision) and WP13 (Coordination of common standards). As future work activity, the matrix of services and solutions, as well as the list of projects and initiatives mentioned above, could be published on a Web page of the APARSEN project and transformed into a navigable public search tool which could be collaboratively extended in the course of the project and even after its conclusion. The key entities identified in this work could also be used to build an ontology representing the relationships and interdependencies between the main interoperability concepts of the digital preservation interoperability landscape (see Figure 15 for an example of relationships between different types of interoperability concepts: interoperability initiative, solution, layer, objects, DP community). The ontology could be made publicly available and maintained as a conceptual long-term contribution of the VCoE to the digital preservation community.

These tools could represent valuable instruments for the VCoE both for raising a common awareness and understanding about interoperability opportunities and challenges in digital preservation and for defining a digital preservation agenda ensuring a coordinated and interoperable digital preservation ecosystem.

7 REFERENCES

ALCTS/CCS Commitee on Cataloging: Description and Access. (2000). *Task Force on Metadata: Final report, June 16, 2000.*

Bazzanella, B., Bortoli, S., & Bouquet, P. (2013). Can Persistent Identifiers be Cool? *Proceedings of the 8th International Digital Curation Conference.* Amsterdam: In press.

Berners-Lee, T. (2001). The Semantic Web. *Scientific American Magazine*.

Berners-Lee, T. (2006, 07 27). *http://www.w3.org/DesignIssues/LinkedData.html.* Tratto da Design Issues.

Bouquet, P., Bazzanella, B., Riestra, R., & Dow, M. (2011). *DIGOIDUNA: Digitla Object Identifiers and Unique Author Identifiers to enable services for data quality assesment, provenance and access.* Tratto da DIGOIDUNA, Digital Object Identifiers and Author Identifiers: http://digoiduna.files.wordpress.com/2011/12/digoiduna_final_report_expert_feedback1.pdf

CabinetOffice. (2004). *e-Government Interoperability Framework.*

CINI, FORTH, SBA. (2012). *D24.1. Report on authenticity and plan for interoperable authenticity evaluation system.*

D'Elia, S., Marchetti, P. G., Coene, Y., & Smolders, S. (2008). Architecture and Services for Computational Intelligence in Remote Sensing. *Studies in Computational Intelligence, 133*, 79-123.

D'Elia, S. (2004). Image information mining for Earth observation at ESA. *Geoscience and Remote Sensing Symposium, IGARSS '04.*

Doerr, M., & Tzitzikas, Y. (2012). Information Carriers and Identification of Information Objects: An Ontological Approach. *CoRR*.

Gentil-Beccot, A., Mele, S., Holtkamp, A., O'Connell, H., & Brooks, T. (2009). Information resources in High-Energy Physics: Surveying the present landscape and charting the future course. *Journal of the American Society for Information Science and technology, 60*(1).

Geraci, A. (1991). *IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries.* IEEE Press.

Giaretta, D. (2007). The CASPAR Approach to Digital Preservation. *The International Journal of Digital Curation, 2*(1).

Giaretta, D. (2011). *Advanced Digital Preservation.* Springer-Verlag Gmbh.

Hedstrom, M. (2001). Exploring the Concept of Temporal Interoperability as a Framework for Digital Preservation. *Third DELOS Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries.* Darmstadt, Germany; September 8–9.

Heery, R., Gardner, T., Day, M., & Patel, M. (2000). *DESIRE metadata registry framework, Deliverable 3.5.*

Heiler, S. (1995). Semantic interoperability. *ACM Computing Surveys (CSUR) , 27*(2).

Hyvönen, E. (2012). Publishing and Using Cultural Heritage Linked Data on the Semantic Web. In *Synthesis Lectures on the Semantic Web: Theory and Technology* (Vol. 2). Palo Alto, CA: Morgan & Claypool .

Lunghi, M., Bellini, E., & Cirinnà, C. (2012). Trust and persistence for Internet resources – Online first: Global Interoperability and Linked Data in Libraries. *Italian Journal of Library and Information Science*.

Miller, P. (2000). *ARIADNE.* Tratto da ARIADNE Web Magazine for Information Professionals: http://www.ariadne.ac.uk/issue24/interoperability

National Information Standards Organization. (2004). *Understanding metadata.* Tratto da http://www.niso.org: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Park, J., & Ram, S. (2004). Information systems interoperability: What lies beneath? *ACM Transactions on Information Systems , 22*(4).

Paskin, N. (2006). Identifier Interoperability - A Report on Two Recent ISO Activities. *D-Lib Magazine, 12*(4).

Rothenberg, J. (2008). *Interoperability as a semantic cross-cutting concern*. Tratto da Interoperabiliteit: Eerlijk zullen we alles delen.

South, D., & Diaconu, C. (2012). The DPHEP Study Group: Data Preservation in High Energy Physics. *ICHEP2012- 36th International Conference for High Energy Physics.* Melburne.

Strodl, S., Petrov, P., & Rauber, A. (2011). *Research on Digital Preservation within projects co-funded by the European Union in the ICT programme.*

Strubulis, C., Tzitzikas, Y., Doerr, M., & Flouri, G. (2012). Evolution of Workflow Provenance Information in the Presence of Custom Inference Rules. *3rd International Workshop on the role of Semantic Web in Provenance Management (SWPM'12), co-located with ESWC'12, Heraklion, Crete, June 2012.*

Taylor, A. G. (2004). The Organization of Information.

Theodoridou, M., Tzitzikas, Y., Doerr, M., & Marketakis, Y. (2010). Modeling and Querying Provenance by Extending CIDOC CRM. *J. Distributed and Parallel Databases (Special Issue Call for Papers: Provenance in Scientific Databases), 27*(2).

Tzitzikas, Y., Marketakis, Y., & Kargakis, Y. (2012). Conversion and Emulation-aware Dependency Reasoning for Curation Services . *Proceedings of the 9th Annual International Conference on Digital Preservation (iPres2012).* Toronto.

Usländer, T., Coene, Y., & Marchetti, P. G. (2012). *Heterogeneous Mission Accessibility Cookbook.* ESA Communications.

van der Hoeven, J. (2010). *Insight into digital preservation of research output in Europe.*

7.1 WEB REFERENCES

CSW-ebRIM	http://www.opengeospatial.org/standards/specifications/catalog
EkoLAB	http://ekolab.iia.cnr.it/home_ENG.htm
GEMET	http://www.eionet.europa.eu/gemet/about
GMES	www.gmes.info/
GEOSS	http://www.earthobservations.org/geoss.shtml
GSCDA	http://gmesdata.esa.int/web/gsc/home
HMA	http://wiki.services.eoportal.org/tiki-index.php?page=HMA%20Wiki
INSPIRE	http://inspire.jrc.ec.europa.eu/
InterRisk	http://interrisk.nersc.no
ISO	www.iso.org/
KEO	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=KEO+Project
KLAUS	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=KLAUS+Project
LTDP	http://earth.esa.int/gscb/objectivesLTDP.html
OTE	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=OTE+Project
OTEG	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=OTEG+Project
RARE	http://rssportal.esa.int/deepenandlearn/tiki-index.php?page=RARE+Project
SMAAD	http://wiki.services.eoportal.org/tiki-index.php?page=SMAAD
W3C	www.w3.org/

8 ACRONYMS

CSW-ebRIM	Catalog Service for the Web - ebRIM
EkoLab	Environmental Knowledge Organisation Laboratory
ENVISION	European NoVel Imaging Systems for ION therapy
EO	Earth Observation
GEMET	GEneral Multilingual Environmental Thesaurus
GEOSS	Global Earth Observation System of Systems
GMES	Global Monitoring for Environment and Security
GSCDA	GMES Space Component Data Access
HMA	Heterogeneous Missions Access
INSPIRE	Infrastructure for Spatial Information in the European Community
ISO	International Standard Organisation
KLAUS	KEO Demonstrator with Models for Land Use Management

KEO	Knowledge-based Earth Observation
LTDP	Long Term Data Preservation
OTE	Ontology and Terminology for Earth Observation
OTEG	Open Access Ontology / Terminology for the GMES Space Component
OWL	Web Ontology Language
RARE	Rapid Response Support Server
RDF	Resource Description Framework
SEPR	Semantic rules for automated Earth Observation products identification
SKOS	Simple Knowledge Organization System
SMAAD	Semantic-Web for Mediated Access Across Domains
SWING	Semantic Web Services Interoperability for Geospatial Decision
W3C	World Wide Web Consortium

Appendix I: Linked Data Library Dataset

BibBase

BibBase.org facilitates the dissemination of scientific publications over the Internet.

British National Bibliography (BNB)

British National Bibliography (BNB) published as Linked Data, linked to external sources including VIAF, LCSH, Lexvo, GeoNames, MARC country, and language, Dewey.info, RDF Book Mashup....

Calames

Calames is the French academic union catalog of archives and manuscripts, maintained by ABES.

Chronicling America

Chronicling America provides access to information about historic newspapers and select digitized newspaper pages. It contains 140,000 newspapers and 3.2 million pages.

Cambridge University Library dataset #1

This data marks the first major output of the COMET project, a JISC-funded collaboration between Cambridge University Library and CARET.

data.bnf.fr - Bibliothèque nationale de France

data.bnf.fr gathers data from the different databases of the Bibliothèque nationale de France, so as to create Web pages about works and authors, together with a RDF view on the extracted data.

Datos.bne.es

The dataset makes available the authority and bibliography catalogue from the Biblioteca Nacional de España (BNE, National Library of Spain) as Linked Data

Scottish Mountaineering Council Journals Issues 1-36

Digital archive of 'Scottish Mountaineering Club Journal Issues 1 to 36, 1890-1901', which was created by Alan Dawson, University of Strathclyde, with funding from the Scottish Mountaineering Trust.

CrossRef DOI Resolver

Digital Object Idenfiers (DOI) are a persistent identifier strategy used by around 3,000 publishers to identify their documents, mostly scholarly publications.

Europeana Linked Open Data

The data.europeana.eu pilot is part of Europeana's effort of making its metadata available as Linked Open Data on the Web. It currently serves metadata on 3.5 million items.

Freebase

Freebase is an open database of the world's information. It is built by the community and for the community — free for anyone to query, contribute to, built applications on top, or integrate into their websites.

Hungarian National Library (NSZL) catalog

OPAC and Digital Library and the corresponding authority data as Linked Open Data.

Linked Periodicals Database

The Linked Periodicals Database is a dataset from the Data Incubator which aggregates journal metadata provided by CrossRef, Highwire Press, and the National Library of Medicine.

lobid. Index of libraries and related organisations lobid-organisations provides URIs for library organizations, based on the existing and well-established International Standard Identifier for Libraries and Related Organizations (ISIL).

lobid. Bibliographic Resources

lobid-resources is a service which offers access to metadata about bibliographic resources (books, articles, pdfs etc.). Currently there are more than 7 million records.

medline

RDF representation of the Medline catalog. Information about 19 million articles linked to http://dx.doi.org/ with article identifiers and http://crossref.org/ with journal identifiers.

NTNU special collections

The digitized historical manuscripts held in the special collections of the Norwegian University of Science and Technology (NTNU).

The Open Library

One Web page for every book ever published. Currently gathers over 20 million records from a variety of large catalogs as well as single contributions.

English Language Books listed in Printed Book Auction Catalogs from 17th Century Holland

The books are those listed in the English-language section of Dutch printed book auction catalogs of collections of scholars and clergy.

ePrints3 Institutional Archive Collection (RKBExplorer)

Linked Data version of a number of ePrints3 archives.

ECS Southampton EPrints

This is live data produced by the EPrints server, which is distinct from the service provided by RKB Explorer.

Sudoc bibliographic data

Sudoc is the French academic union catalog, maintained by ABES. It contains 10 million bibliographic records.

Open Library data mirror in the Talis Platform

Modeled using the JSON data dumps from the Open Library. Provides a SPARQL endpoint and OpenSearch interface (with RSS 1.0 output).

theses.fr

theses.fr is the french dissertations search engine, maintained by ABES.

Linked Data Service der Universitätsbibliothek Mannheim

Publishes RDF for a number of bibliographic resources: Bibliograhic data of the Südwestdeutscher Bibliotheksverbund, Bibliographic data of the Hessisches Bibliotheksinformationssystem, and others.

University of Sussex Reading Lists

Linked Data version of the resources available through the university's reading-list search engine.

20th Century Press Archives

More than 30 million documents, mostly press clippings about individual persons, companies and other corporate bodies, products and a wide variety of economics-related topics.